UNIVERSIDADE DE LISBOA
FACULDADE DE MEDICINA VETERINÁRIA

DEVELOPMENT OF A NOVEL PLATFORM FOR HIGH-THROUGHPUT GENE DESIGN AND ARTIFICIAL GENE SYNTHESIS TO PRODUCE LARGE LIBRARIES OF RECOMBINANT VENOM PEPTIDES FOR DRUG DISCOVERY

ANA FILIPA PEREIRA SEQUEIRA

Tese especialmente elaborada para obtenção do grau de Doutor em Ciências Veterinárias na Especialidade de Ciências Biológicas e Biomédicas

Orientadores:
Doutor Carlos Mendes Godinho de Andrade Fontes
Doutor Renaud Vincentelli
Doutora Catarina Isabel Proença Duarte Guerreiro

2016

UNIVERSIDADE DE LISBOA
FACULDADE DE MEDICINA VETERINÁRIA

DEVELOPMENT OF A NOVEL PLATFORM FOR HIGH-THROUGHPUT GENE DESIGN AND ARTIFICIAL GENE SYNTHESIS TO PRODUCE LARGE LIBRARIES OF RECOMBINANT VENOM PEPTIDES FOR DRUG DISCOVERY

ANA FILIPA PEREIRA SEQUEIRA

Tese especialmente elaborada para obtenção do grau de Doutor em Ciências Veterinárias na Especialidade de Ciências Biológicas e Biomédicas

Orientadores:
Doutor Carlos Mendes Godinho de Andrade Fontes
Doutor Renaud Vincentelli
Doutora Catarina Isabel Proença Duarte Guerreiro

Constituição do júri:
Presidente:
Doutor Rui Manuel de Vasconcelos e Horta Caldeira
Vogais:
Duarte Miguel de França Teixeira dos Prazeres
Doutor José António Mestre Prates
Doutor Carlos Mendes Godinho de Andrade Fontes
Doutora Lucília Maria Alves Ribeiro Domingues
Doutor Renaud Vincentelli
Doutor Pedro Miguel Dias Castanheira

2016

À minha Família

Agradecimentos

Um doutoramento é sem dúvida uma longa viagem repleta de momentos altos e baixos em busca de novo conhecimento científico. Passados quatro anos, reconheço em mim muitas mudanças, quer pessoais quer profissionais, uma vez que aprendi muito, assimilei novo conhecimento cientifico e cresci muito como investigadora. Todo este esforço e trabalho culminam agora nesta tese de doutoramento que pretende reflectir toda a minha aprendizagem e dedicação ao longo destes anos. Contudo, o sucesso desta etapa não seria possível sem ajuda de muitas pessoas que sempre se mostraram disponíveis para me ajudar. Na verdade, esta aventura foi vivida ao lado das pessoas certas, muitas das quais encontrei na NZYTech que tão bem que me acolheu. Assim, quero agradecer a participação de várias pessoas e entidades no desenvolvimento deste trabalho. A todos, o meu muito sincero obrigado:

À Fundação para a Ciência e Tecnologia por parte do financiamento da minha bolsa de doutoramento;

À NZYTech, por me ter acolhido como estudante de doutoramento em empresa e financiado parte da minha bolsa. Pelos meios físicos e materiais que me disponibilizou para a realização de todo o trabalho. Por me proporcionar uma experiência incrível em ambiente empresarial e permitir que faça parte deste projecto de vida;

À Faculdade de Medicina Veterinária e ao CIISA por me terem aceite como estudante de doutoramento e por terem disponibilizado meios físicos e materiais para a realização de todo o trabalho;

Ao Professor Doutor Carlos Fontes pelo excepcional orientador que é e por ser um exemplo de como deve ser um investigador de alto nível. Por todos os ensinamentos que me transmitiu, pela paciência e pela orientação. Agradeço o seu entusiasmo constante, boa disposição e amizade. Pela oportunidade de integrar esta equipa e por poder contribuir para o crescimento da nossa empresa. Obrigada por acreditar em mim e por me fazer crescer como investigadora.

To Renaud Vincentelli, my co-supervisor, for his support during my doctoral project. Thank you for your helpful advices and for the careful revision of this thesis;

To each one of VENOMICS partners whose contribution made this work possible. I would like to thank for the opportunity to meet wonderful and professional people, and to embark on this amazing project;

Ao Professor Doutor Luís Ferreira pela sua orientação e amizade. Pelas sua palavras sábias nos momentos certos, pelo entusiasmo e pela sua simpatia;

Ao Professor Doutor José Prates pela sua simpatia e disponibilidade, e pelos seus ensinamentos na área da Bioquímica;

À Patrícia Ponte e à Catarina Guerreiro por me terem acolhido tão bem neste grupo. Por serem os pilares desta empresa e pela sua capacidade excepcional de trabalho. Obrigada pela simpatia, disponibilidade, entreajuda e amizade. Em particular, agradeço à Catarina pela sua disponibilidade e revisão cuidada da minha tese. E também pelos bons momentos que passámos juntas durante as viagens do VENOMICS;

À Joana Brás e à Vânia Fernandes pela ajuda constante e pela partilha de bons e maus resultados experimentais. Obrigada pelas nossas discussões científicas, pelos conselhos e pelo vosso espírito crítico. Pela amizade e pela boa disposição. Em particular, à Vânia Fernandes, por termos partilhado a aventura de fazer doutoramento em empresa e por termos vivido descobertas, frustrações e alegrias juntas. À Joana pelo companheirismo, empatia e pela excelente capacidade de trabalho. Por termos produzido milhares de genes juntas;

À equipa comercial, André Fernandes, Diogo Comprido e Pedro Rio, pela boa disposição e pela ajuda sempre presente. Pela amizade e pelos almoços exóticos para a partilha de ideias.

À Andreia Peixoto, ao João Belchior, à Nádia Rodrigues, à Vânia Cardoso e à Maria João pela boa disposição e sentido de humor. Pela simpatia e pela capacidade de trabalho que fazem com que a NZYTech cresça cada vez mais. Em particular, agradeço à Andreia pelos bons momentos, pelo companheirismo e pela partilha de gostos em comum;

A todos os meus amigos que me têm acompanhado nesta etapa, pela amizade, pelo apoio e pelos bons momentos que temos partilhado. Em particular, agradeço à Paula, à Mafalda e à Helena pela amizade e carinho, por acreditarem nas minhas capacidades e por estarem sempre ao meu lado;

À Inês Sousa pela sua amizade, disponibilidade e pela excelente cientista que é. Por ter acompanhado o meu crescimento como investigadora e por me motivar sempre a alcançar os meus objectivos. Obrigada por todo o apoio e por seres uma excelente pessoa;

À minha irmã, Mónica, agradeço por ter sido um exemplo para mim. Pela amizade, pelos conselhos, pela paciência, pela cumplicidade e pelos bons momentos que vivemos. Agradeço-te por sempre acreditares em mim e pela motivação constante;

Aos meus pais, Maria Virgínia e Etelvino, pelo amor e carinho, pelo apoio e por estarem sempre presentes. Sem vocês não teria chegado aqui e espero partilhar convosco muitas mais conquistas da minha vida;

Ao Rodrigo por toda ajuda e apoio, por me ouvires inúmeras vezes a falar de ciência e por partilharmos este gosto. Pela paciência sem fim e pela motivação. Pelo amor incondicional, pela amizade e pelo projecto lindo de vida que estamos a construir, e que nos enche o coração todos os dias.

This work was co-funded by Fundação para a Ciência e a Tecnologia, grant SFRH/BDE/51602/2011 from Ministério da Ciência, Tecnologia e Ensino Superior, and by NZYTech, Lda, genes & enzymes

RESUMO

Desenvolvimento de uma nova plataforma de alta capacidade para desenhar e sintetizar genes artificiais, para a produção de péptidos venómicos recombinantes

Os venenos animais são misturas complexas de moléculas biologicamente activas que se ligam com elevada selectividade e eficácia a uma grande variedade de receptores de membrana. Embora apresentem baixa imunogenicidade, os venenos podem afectar a função celular actuando ao nível dos seus receptores. Actualmente, pensa-se que os venenos de animais constituam uma biblioteca natural de mais de 40 milhões de moléculas diferentes que têm sido continuamente aperfeiçoadas ao longo do processo evolutivo. Tendo em conta a composição dos venenos, os péptidos reticulados são a classe mais atractiva de moléculas com interesse farmacológico. No entanto, a utilização de venenos para o desenvolvimento de novos fármacos está limitada por dificuldades em obter estas moléculas em quantidades adequadas ao seu estudo. Neste trabalho desenvolveu-se uma plataforma de alta capacidade para a síntese de genes sintéticos codificadores de péptidos venómicos, com o objectivo de produzir bibliotecas de péptidos venómicos recombinantes que possam ser rastreadas para a descoberta de novos medicamentos. Com o objectivo de sintetizar genes pequenos (< 500 pares de bases) com elevada fidelidade e em simultâneo, desenvolveu-se uma metodologia de PCR (polymerase chain reaction) robusta e eficiente, que se baseia na extensão de oligonucleótidos sobrepostos. Para possibilitar a automatização da plataforma de síntese de genes, foram construídas duas ferramentas bioinformáticas para desenhar simultaneamente dezenas a milhares de genes optimizados para a expressão em *Escherichia coli* (ATGenium) e os respectivos oligonucleótios sobrepostos (NZYOligo designer). Esta plataforma foi optimizada para sintetizar em simultâneo 96 genes sintéticos, tendo-se obtido uma taxa de erro de 1.1 mutações por kb de DNA sintetizado. A fim de diminuir a taxa de erro associada à produção de genes sintéticos, desenvolveu-se um método para remoção de erros utilizando a enzima T7 endonuclease I. A enzima T7 endonuclease I mostrou-se muito eficaz no reconhecimento e clivagem de moléculas DNA que apresentam emparelhamentos incorrectos, reduzindo drasticamente a frequência de erros identificados em genes grandes, de 3.45 para 0.43 erros por kb de DNA sintetizado. Investigou-se também a influência do desenho dos genes, da presença de tags de fusão, da localização celular da expressão e da actividade da protease *Tobacco Etch Virus* (TEV) para a remoção eficiente de tags, na expressão de péptidos venómicos ricos em cisteínas em *E. coli*. A utilização de codões meticulosamente escolhidos afectou drasticamente os níveis de expressão em *E. coli*. Para além disso, os resultados mostram que existe uma pressão significativa no uso dos dois codões que codificam para o resíduo cisteína, o que sugere que ambos os codões têm de estar presentes, em níveis equivalentes, nos genes que foram desenhados e optimizados para garantir elevados níveis de expressão. Este trabalho indicou também que o tag de fusão DsbC foi o mais apropriado para a expressão eficiente de péptidos venómicos ricos em cisteínas, particularmente quando os péptidos recombinantes foram expressos no periplasma bacteriano. Confirmou-se que a protease TEV é eficaz na remoção de tags de fusão, podendo o seu local de reconhecimento conter quaisquer aminoácidos na extremidade C-terminal, com excepção da prolina. Desta forma, verificou-se não ser necessário incorporar qualquer aminoácido extra na extremidade N-terminal dos péptidos venómicos recombinantes. Reunindo todos os resultados, verificou-se que a *E. coli* é um hospedeiro adequado para a expressão, na forma solúvel, de péptidos venómicos potencialmente funcionais. Por último, foram produzidos, com uma taxa de erro reduzida, ~5000 genes sintéticos codificadores de péptidos venómicos utilizando a nova plataforma de elevada capacidade para a síntese de genes aqui desenvolvida. A nova biblioteca de genes sintéticos foi usada para produzir a maior biblioteca de péptidos venómicos recombinantes construída até agora, que inclui 2736 péptidos venómicos. Esta biblioteca recombinante está presentemente a ser rastreada com o objectivo de descobrir novas drogas com interesse para a saúde humana.

Palavras-chave: péptidos venómicos recombinantes, técnica de alta capacidade para a síntese de genes, utilização de codões, métodos para a remoção de erros, protease TEV

ABSTRACT

Development of a novel platform for high-throughput gene design and artificial gene synthesis to produce large libraries of recombinant venom peptides for drug discovery

Animal venoms are complex mixtures of biologically active molecules that, while presenting low immunogenicity, target with high selectivity and efficacy a variety of membrane receptors. It is believed that animal venoms comprise a natural library of more than 40 million different natural compounds that have been continuously fine-tuned during the evolutionary process to disturb cellular function. Within animal venoms, reticulated peptides are the most attractive class of molecules for drug discovery. However, the use of animal venoms to develop novel pharmacological compounds is still hampered by difficulties in obtaining these low molecular mass cysteine-rich polypeptides in sufficient amounts. Here, a high-throughput gene synthesis platform was developed to produce synthetic genes encoding venom peptides. The final goal of this project is the production of large libraries of recombinant venom peptides that can be screened for drug discovery. A robust and efficient Polymerase Chain Reaction (PCR) methodology was refined to assemble overlapping oligonucleotides into small artificial genes (< 500 bp) with high-fidelity. In addition, two bioinformatics tools were constructed to design multiple optimized genes (ATGenium) and overlapping oligonucleotides (NZYOligo designer), in order to allow automation of the high-throughput gene synthesis platform. The platform can assemble 96 synthetic genes encoding venom peptides simultaneously, with an error rate of 1.1 mutations per kb. To decrease the error rate associated with artificial gene synthesis, an error removal step using phage T7 endonuclease I was designed and integrated into the gene synthesis methodology. T7 endonuclease I was shown to be highly effective to specifically recognize and cleave DNA mismatches allowing a dramatically reduction of error frequency in large synthetic genes, from 3.45 to 0.43 errors per kb. Combining the knowledge acquired in the initial stages of the work, a comprehensive study was performed to investigate the influence of gene design, presence of fusion tags, cellular localization of expression, and usage of Tobacco Etch Virus (TEV) protease for tag removal, on the recombinant expression of disulfide-rich venom peptides in *Escherichia coli*. Codon usage dramatically affected the levels of recombinant expression in *E. coli*. In addition, a significant pressure in the usage of the two cysteine codons suggests that both need to be present at equivalent levels in genes designed *de novo* to ensure high levels of expression. This study also revealed that DsbC was the best fusion tag for recombinant expression of disulfide-rich peptides, in particular when expression of the fusion peptide was directed to the bacterial periplasm. TEV protease was highly effective for efficient tag removal and its recognition sites can tolerate all residues at its C-terminal, with exception of proline, confirming that no extra residues need to be incorporated at the N-terminus of recombinant venom peptides. This study revealed that *E. coli* is a convenient heterologous host for the expression of soluble and potentially functional venom peptides. Thus, this novel high-throughput gene synthesis platform was used to produce ~5,000 synthetic genes with a low error rate. This genetic library supported the production of the largest library of recombinant venom peptides constructed until now. The library contains 2736 animal venom peptides and it is presently being screened for the discovery of novel drug leads related to different diseases.

Key-words: recombinant venom peptides, high-throughput gene synthesis, codon usage, error removal, TEV cleavage

This thesis was based on the following manuscripts:

I. **Sequeira, A.F.**, Brás, J.L.A., Guerreiro, C.I.P.D., Vincentelli, R., Fontes, C.M.G.A. (2016) Development of a gene synthesis platform for the efficient large scale production of small genes encoding animal toxins. (Manuscript submitted)

II. **Sequeira, A.F.**, Guerreiro, C.I.P.D., Vincentelli, R., Fontes, C.M.G.A. (2016) T7 endonuclease I mediates error correction in artificial gene synthesis. *Molecular Biotechnology*, 58(8-9), 573-84.

III. **Sequeira, A.F.***, Turchetto, J.*, Saez, N.J., Peysson, F., Ramond, L., Duhoo, Y., Blémont, M., Fernandes, V.O., Gama, L.T., Ferreira, L.M.A., Guerreiro, C.I.P.D., Darbon, H., Fontes, C.M.G.A., Vincentelli, R. (2016) Gene design, fusion technology and TEV cleavage site influence the expression of disulfide-rich venom peptides in *Escherichia coli*. (Accepted in Microbial Cell Factories)

IV. Turchetto, J.*, **Sequeira, A.F.***, Ramond, L. *, Peysson. F. *, Brás, J. L.A., Saez, N.J., Duhoo, Y., Blémont, M., Guerreiro, C.I.P.D., Gilles, N., Darbon, H., Fontes, C.M.G.A., Vincentelli, R. (2016) High-throughput expression of animal venom toxins in *Escherichia coli* to generate a large library of recombinant reticulated peptides for drug discovery. (Accepted in Microbial Cell Factories)

* Authors contributed equally to the work

The Author's contribution:

Publication I: Collaborated in experimental design. Performed all molecular biology experiments designed to develop the high-throughput gene synthesis platform. Was involved in the development of two bioinformatics tools. Wrote the manuscript.

Publication II: Collaborated in experimental design. Performed all molecular biology experiments. Constructed the error correction assay. Wrote the manuscript.

Publication III: Collaborated in experimental design. Was involved in gene synthesis, cloning, recombinant protein expression of venom peptides. Constructed some of the expression pHTP-derivative vectors. Major writing contributions.

Publication IV: Collaborated in experimental design. Was involved in the large scale gene design, synthesis and cloning of genes encoding venom peptides. Major writing contributions.

INDEX

LIST OF TABLES

LIST OF FIGURES

LIST OF ABBREVIATIONS AND SYMBOLS

%	Percentage
e^-	electron
Å	Angstrom
A	Adenine
aa	Amino acid
ACN	Acetonitrile
ANOVA	Analysis of variance
Arg	Arginine
Asn	Asparagine
AT	Adenine-thymine
ATDB	Animal Toxin Database
ATG	Start codon
ATP	Adenosine Triphosphate
BL21(DE3)	*E. coli* expression strain containing the DE3 lysogen that carries the gene for T7 RNA polymerase under control of the *lac*UV5 promoter
BL21(DE3) pLysS	*E. coli* expression strain which carries both the DE3 lysogen and the plasmid pLysS, which reduces the basal expression of recombinant genes by inhibiting the basal levels of T7 RNA polymerase.
BLAST	Basic Local Alignment Search Tool
bp	base-pair
C	Cytosine
$CaCl_2$	Calcium chloride
CAI	Codon Adaptation Index
cAMP	Cyclic adenosine monophosphate
CAP	Catabolite activator protein
cat	Chloramphenicol-acethyltransferase gene
cDNA	Coding Desoxyribonucleic Acid
cm	Centimeter
Cys	Cysteine
CPG	Controlled Pore Glass
CRISPR	Clustered regularly-interspaced short palindromic repeats
DA-PCR	Dual Asymetric-Polymerase Chain Reaction
DMT	Dimethoxytrityl
DNA	Desoxyribonucleic Acid
dNTP	Desoxynucleotide
ds	double strand
DsbA	Disulfide oxireductase A
DsbB	Disulfide oxireductase B
DsbC	Disulfide isomerase C
DsbD	Disulfide oxireductase D
DTT	Dithiothreitol
E. coli	*Escherichia coli*
EDTA	Ethylenediamine tetraacetic acid
EMC	Enzymatic Mismatch Cleavage
ER	Endoplasmic reticulum

ESI	Electrospray Ionisation
EST	Expressed Sequence Tag
FDA	Food and Drug Administration
Fh8	Calcium binding protein Fh8
G	Guanine
GC	Guanine-cytosine
GFP	Green Fluorecent Protein
Glu	Glutamate
Gly	Glycine
GPCRs	G-protein coupled receptors
GST	Gluthathione S-Transferase
h	hour
H	Fusion tag
HCl	Hydrogen chloride
HEPES	2-[4-(2-hydroxyethyl)-1-piperazinyl] ethanesulfonic acid
HGP	Human Genome Project
His	Histidine (H)
His_6/6HIS	Hexa-histidine tag
HPLC	High Performance Liquid Chromatography
HTP	High-throughput
HTS	High-throughput screens
IB	Inclusion Bodies
Ile	Isoleucine
IMAC	Immobilized Metal Affinity Chromatography
IPS	Improved PCR Synthesis
IPTG	Isopropyl β-D-1-thiogalactopyranoside
Kan	Kanamycin
kb	kilobase
kDa	kiloDalton
Kg	Kilogram
L	Litre
lac	Lactose gene
LacI	Lactose repressor
lacO	Lactose operator gene
LB	Luria Bertani
LC	Liquid Chromatography
LCR	Ligase Chain Reaction
LIC	Ligation-independent cloning
M	Molar
MALDI-TOF	Matrix-Assisted Laser Desorption/Ionisation Time-of-Flight
MBP	Maltose Binding Portein
mg	Milligram
$MgCl_2$	Magnesium chloride
min	Minutes
mL	Milliliter
mm	Milimeter
mM	Millimolar
mRNA	Messenger Ribonucleic Acid
MS	Mass spectrometry

MS/MS	Tandem mass spectrometry
MutH	DNA mismatch repair protein MutH
MutL	DNA mismatch repair protein MutL
MutS	DNA mismatch repair protein MutS
MW	Molecular Weight
N	Nitrogen
NaCl	Sodium chloride
NaOH	Sodium hydroxide
NET	Noradrenaline transporter
ng	nanogram
NGS	Next generation sequencing
Ni^{2+}	Nickel ion
nm	Nanometer
nM	Nanomolar
NMDA	N-methyl-D-aspartate
nt	Nucleotide
NusA	N-utilization substance protein A
ºC	Celcius degree
OE-PCR	Overlap Extension-Polymerase Chain Reaction
OH	Hydroxil group
ORF	Open Reading Frame
O_2	Oxygen
PAGE	Polyacrylamide Gel Electrophoresis
PAS	PCR-based Accurate Synthesis
PCA	Polymerase Chain Assembly
PCA-DT	Polymerase Chain Assembly using DNA template
PCA-DTF	Polymerase Chain Assembly DNA template-free
PCR	Polymerase Chain Reaction
PDI	Protein Disulfide Isomerase
pH	negative decimal logarithm of the hydrogen ion activity in a solution
Phe	Phenilalanine
PLA_2	Phospholipases A_2
pLysS	Plasmid for expressing T7 lysozyme
pmol	picomole
PS	Polystyrene
PTDS	PCR-based Two-steps DNA Synthesis
PTM	Post-translational modification
RBS	Ribosomal Binding Site
RNA	Ribonucleic Acid
rpm	Rotations per minute
rRNA	Ribosomal Ribonucleic Acid
s	second
SAS	Statistical Analysis Software
SD	Shine-Dalgarno
SDS	Sodium Dodecyl Sulfate
Ser	Serine
SGS	Simplified Gene Synthesis
SLIC	Sequence- and ligation-independent cloning
SPE	Solid Phase Extraction

SPPS	Solid phase peptide synthesis
sss	single-strand specific
SUMO	Small ubiquitin-like modifier
T	Thymine
TALE	Transcription activator-like effector
TBIO	Thermodynamically Balanced Inside-Out
TEV	Tobacco etch virus
TEV_{SH}	Mutant of TEV
Tm	Melting temperature
Tris	2-Amino-2-hydroxymethyl-propane-1,3-diol
tRNA	Transfer Ribonucleic Acid
Trx	Thioredoxin
U	Enzymatic units
Ub	Ubiquitin
UHPLC-MS	Ultra-High Performance Liquid Chromatography-Mass Spectrometry
UV	Ultraviolet light
VGCs	Voltage-gated ion channels
v/v	Volume per volume
w/v	Weight per volume
w/w	Weight per weight
WGS	Whole-genome shotgun
$ZnCl_2$	Zinc chloride
µg	Microgram
µL	Microliter
µM	Micromolar
µm	Micrometer
16S rRNA	16S subunit of ribosomal Ribonucleic Acid

1. INTRODUCTION AND THESIS OUTLINE

Today the pharmaceutical industry faces the challenge of rapidly finding new and innovative drugs. Venomous animals express a complex battery of reticulated peptides that present structural and pharmacological diversity. Venoms represent a mostly unexplored reservoir of bioactive peptides that comprise millions of different molecules with high potential for drug discovery. These venom peptides have been fine-tuned during the course of evolution to display not only formidable affinity and selectivity for a variety of cell surface receptors, such as ion channels, but also low immunogenicity and high stability. In general, venom peptides are small disulfide-rich molecules with no more than 120 residues and include up to eight disulfide bonds that are critical for both biological activity and stability. Recently, venom peptides have been the subject of intense investigation and their use as innovative drugs was described in several studies. Thus, venom peptides are an important class of potentially novel therapeutics in modern drug discovery pipelines. Unfortunately, the use of venom peptides as therapeutic or biotechnological molecules is still mostly an unrealized prospective due to difficulties in producing, both synthetically or recombinantly, active molecules in sufficient amounts. Another drawback underlying the low capacity to produce recombinant venom peptides is the availability of biological material to obtain genes that encode these toxins. DNA templates are often not readily available due to the limiting amounts of biological material available from small animals.

The technology to write genetic information encoding any protein that needs to be expressed or analysed is emerging as a paradigm change in the field of recombinant protein production. *De novo* gene synthesis has proven to be the best way to acquire target genes encoding peptides identified in animal venoms. The ability to design DNA sequences according to critical parameters, such as applying the host codon usage, has improved the capacity to recombinantly produce functional proteins in heterologous systems. In addition, it is now well recognized that optimal and efficient production of venom peptides requires the development of high-throughput methods for gene synthesis, gene cloning, recombinant protein expression and protein purification for the rapid characterization of peptides, including identification of their pharmacological interest to accelerate drug discovery. In this respect, several high-throughput strategies for cloning and expression of large numbers of genes are being applied in different laboratories.

Despite rapid developments in gene synthesis technology, fast and accurate synthesis of multiple genes simultaneously is still not well established. High error rates and low throughput are the current difficulties described when producing high fidelity DNA sequences encoding the desired recombinant protein. The focus of this thesis was the development of a novel platform for fast and high fidelity synthesis of small synthetic genes, to allow producing recombinant venom peptides with high stability and in sufficient amounts to explore their

pharmacological properties. The aim of this project was not only the development of a highly efficient high-throughput gene synthesis platform that could benefit many biotechnological applications, but also to elucidate several parameters concerning gene design and synthesis, which can improve the recombinant expression of disulfide-rich peptides in *Escherichia coli*. The research work was mostly developed at NZYTech, a Biotech Company based in Lisbon, and was integrated in the activities of the FP7 funded project VENOMICS. This thesis combines fundamental and applied research from a pragmatic perspective and aims to develop novel technologies that can accelerate biotechnology. In this research context, the doctoral project was applied to address specific problems and orientated by fundamental and applied needs and demands. Beside many other contributions, the development of a completely novel high-throughput gene synthesis platform that could be used to synthesise any DNA sequence to be expressed in *E. coli* was the main achievement of this work.

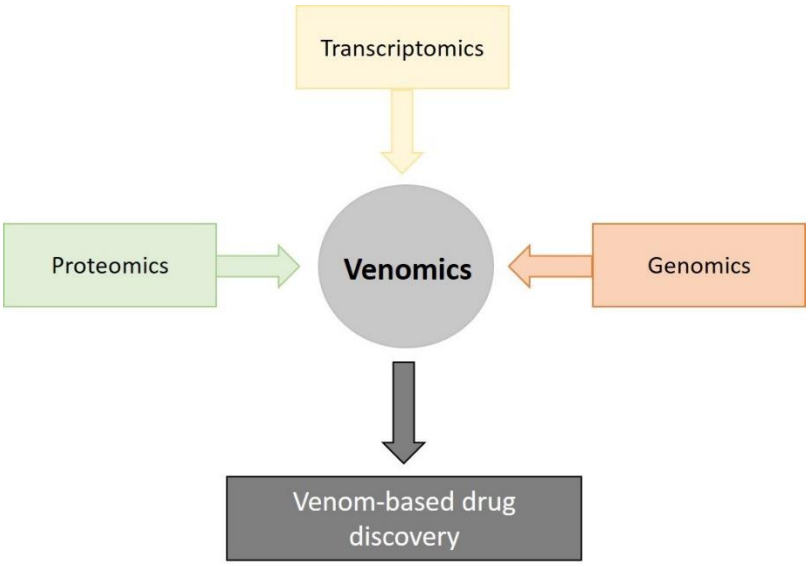
Following this introductory section, this thesis is divided into 7 additional chapters. The second chapter comprises a state of the art review. Several concepts concerning venomous animals and their diversity, with a special focus on the complexity and functionality of venom components, in particular the enormous pharmacological potential of venom peptides components, are revised. In addition, a general description of current gene synthesis methodologies and the impact of high-throughput protocols developed for synthetic biology and also for cloning and protein expression is included. The objectives of this work are specifically defined at the end of the bibliographic review. Chapters 3, 4, 5 and 6 are organized in papers based on scientific manuscripts, in preparation or already submitted to international peer reviewed journals. Finally, the last chapter of the thesis discusses and integrates the results presented in each one of the previous chapters. Future perspectives for the scientific knowledge attained with this work are also approached in this last section.

2. BIBLIOGRAPHIC REVIEW AND OBJECTIVES

2.1. Venomics

The extraordinary potency and pharmacological diversity of animal venoms have made them an increasingly valuable source of lead molecules for drug discovery (Vetter *et al*., 2011). Nevertheless, most of venom chemical diversity remains uncharacterized, in part because of the small quantities available to analysis and also to the low-throughput methodologies used to characterize animal toxin diversity. Similarly to the other “omics”, venom specialists have created a new biological approach not only to understand the complexity of venoms, but also to overcome the experimental and technological bottlenecks recently identified. This approach includes biological methods, such as genomics, transcriptomics and proteomics, and was termed “venomics” (Figure 2.1). The term “venomics” was first introduced in the early 2000s for a description of snake venom composition (Bazaa, Marrakchi, El Ayeb, Sanz, & Calvete, 2005; Juarez, Sanz, & Calvete, 2004) and since then it has been extensively employed in numerous studies.

Figure 2.1| Venomics-based discovery approach.



Venomics use a combination of proteomic, transcriptomic and genomic approaches, and bioinformatic tools to improve venom research strategies.

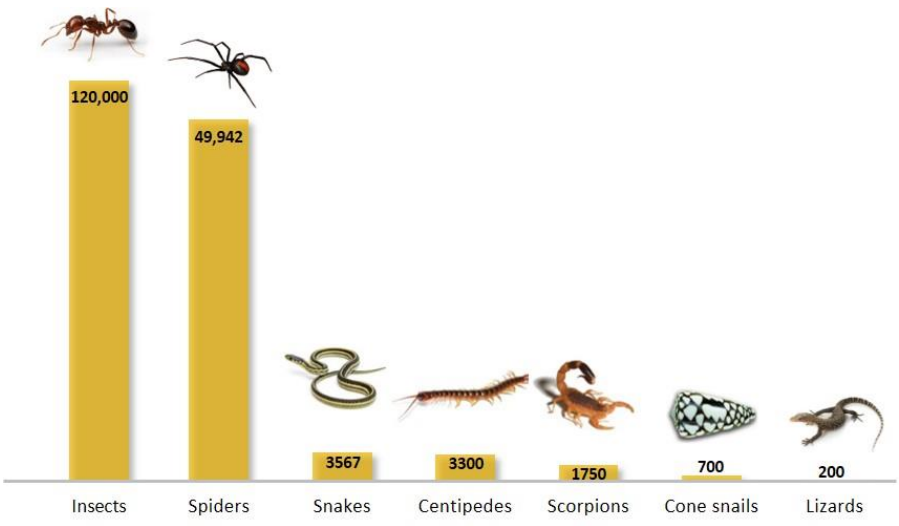
Over the last few years, high-throughput (HTP) technologies are becoming increasingly instrumental in venom research. These allow accelerating the rate of discovery and facilitate the characterization of venom peptides, changing the paradigm of venom exploration from low throughput one gene, one peptide approaches. Combined advances in transcriptomics, like the employment of next generation sequencing (that allows analysing entire transcriptomes of venom glands) and proteomics, through *de novo* sequencing of peptides from small amounts of material, give access to complete information about a single venom. Venomics research is

then nowadays driven by HTP techniques to understand the powerful ability that venomous animals have to prey capture or self-defence, and in what way this process could be helpful to human health (see 2.5 section).

2.1.1. Diversity of venomous animals

Throughout human history, venomous animals have been the subject of considerable public interest, in large part due to the inherent danger associated with their unpleasant effects, the high number of venomous species present around the world and the apparent incongruity between their small and often fragile-looking and the devastating damage they can inflict (Casewell, Wüster, Vonk, Harrison, & Fry, 2013). On a global basis envenomation constitutes a highly relevant public health issue, as there are venomous organisms in every continent and almost every country. However, venomous animals are particularly abundant in tropical regions, which represent the *kitchen of evolution*. Animals that use venoms for defence or predation are widely spread through the animal kingdom, comprising about 170,000 species (Aili *et al*., 2014; Takacs & York, 2014) distributed among all major phyla, such as chordates (reptiles, fishes, amphibians, mammals), echinoderms (starfishes, sea urchins), molluscs (cone snails, octopi), annelids (leeches), nemertines, cnidarians (sea anemones, jellyfish, corals) and the unquestioned champions of the venom world, the arthropods (ants, bees, caterpillars, centipedes, flies, mites, mosquitoes, scorpions, spiders, ticks, reduviid bugs and wasps). Approximately 90% of venomous animal diversity correspond to small animals (< 1 cm long) (Gilles & Servent, 2014), as the large ones such as snakes, tarantulas or blue-ring octopus represent only the smallest part of this diversity. Hymenoptera are the most important venomous order among insects, with more than 120,000 species (Aili *et al.*, 2014), including small animals with size ranging from 0,1 mm to 10 cm long. Other species widely studied for their venom content are distributed by diversified orders and sub-orders (Figure 2.2).

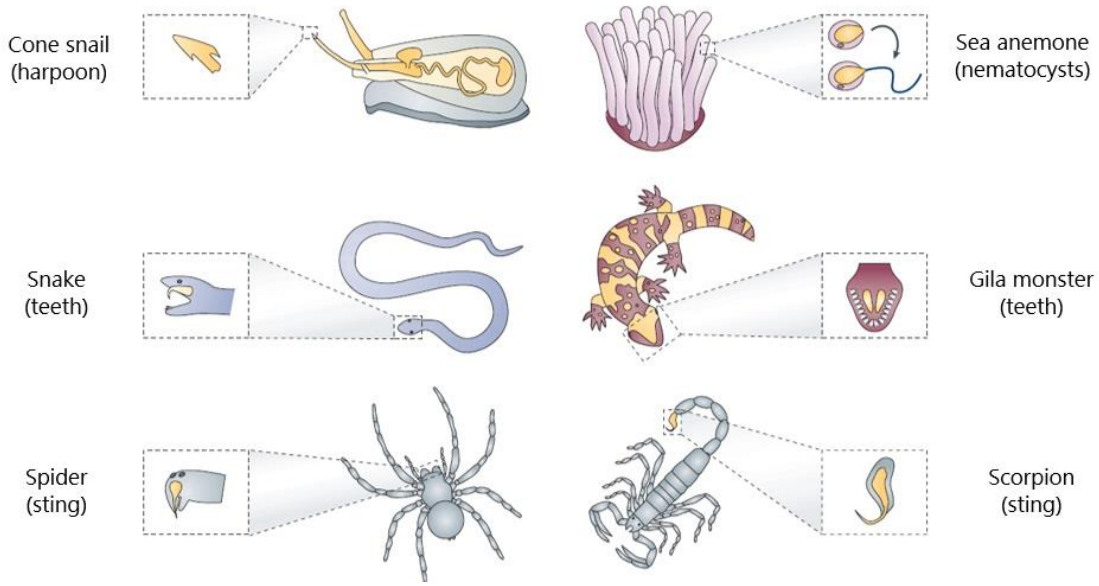
Figure 2.2 | Number of venomous species distributed by diversified orders.



For example, there are 49,942 species of spiders ("World Spider Catalog," 2016), 3567 species of snakes (Uetz & Hošek, 1996), 3300 species of centipedes (King, 2011), 1750 species of scorpions, 700 species of cone snails and 200 species of lizards (King, 2011).

The ability of venom production represents an adaptive trait selected by Natural Selection to act on vital systems of the prey or predator. This ecological advantage allowed that a range of animals evolved to use their venoms for purposes of predation, defence or competitor deterrence. As a result, animal venoms have extremely high specificity and potency for their molecular target due to millions of years of evolutionary fine tuning. Venomous animals developed sophisticated and specialized envenomation systems that have evolved to facilitate the delivery of venoms, and these include fangs, barbs, beaks, modified teeth, harpoons, nematocysts, pincers, proboscises, spines, sprays, spurs and stingers (Figure 2.3) (Casewell *et al.*, 2013; Lewis & Garcia, 2003).

Figure 2.3 | The envenomation apparatus of venomous animals.



Different envenomation systems such as harpoons, nematocysts, teeth and stings are highlighted. Figure adapted from Lewis & Garcia (2003).

Proteomic and transcriptomic analyses have demonstrated that individual spider and cone snail venoms can comprise up to 1000 distinct peptides (Davis, Jones, & Lewis, 2009; Escoubas, Sollod, & King, 2006), while scorpions venoms often contain several hundred components (King, 2011). Therefore, the global animal venom resource can be seen as a collection of more than 40,000,000 biologically active peptides and proteins, if ones uses a conservative estimate of 170,000 species and 250 peptides per venom. For spiders, the most successful terrestrial predators, even using conservative estimates of 50,000 species and 200 peptides/venoms, it could be considered that spider venoms contain at least 10 million bioactive peptides (King, 2011). The total estimated number of bioactive peptides identified in

animal venoms is extraordinary, although only a reduced number of peptides have been characterized (Lewis & Garcia, 2003). Animal diversity highlights the importance of venoms as an unexplored reservoir of multiple bioactive components that can play a prime role in drug discovery.

2.1.2. Animal venoms

A venom can be broadly defined as a 'secretion, produced by a specialised gland in one animal and delivered to a target animal through the infliction of a wound, which contains molecules that disrupt normal physiological or biochemical processes so as to facilitate feeding or defence by the producing animal' (King, 2011). The composition and targeting of venoms reflect their function: fishes or bees produce defensive venoms, being highly conserved and with immediate action. By contrast, predatory venoms are more complex and often variable in composition and physiological effects. Numerous studies have shown that venom components are endowed with remarkable biological properties associated with their capacity to act in a large number of molecular receptors, in the process of incapacitating their target organisms (Casewell *et al*., 2013).

Animal venoms are complex mixtures of hundreds to thousands of bioactive molecules, mainly composed of proteins including enzymes (>10 kDa) and peptides (1-10 kDa), but also containing inorganic salts and small organic molecules (Escoubas & King, 2009; Escoubas, Quinton, & Nicholson, 2008; King, 2011). Peptides are called toxins and are characterized by their stability and their enrichment in disulfide bonds (Gilles & Servent, 2014). In addition, high-molecular-weight proteins mainly have enzymatic activity such as phospholipases A2 (PLA_2), metallo- or serine proteases, L-amino-oxidases and hydrolases. Enzymes are present in a high proportion in snake venoms. Previous studies have suggested that PLA_2 may be important for initiating the digestion of the envenomed prey tissues (Näreoja & Näsman, 2012). In contrast, venoms of cone snails, scorpions and spiders contain a large proportion of reticulated peptides, most of them tightly folded and stabilized by several disulfide bridges (Escoubas & King, 2009).

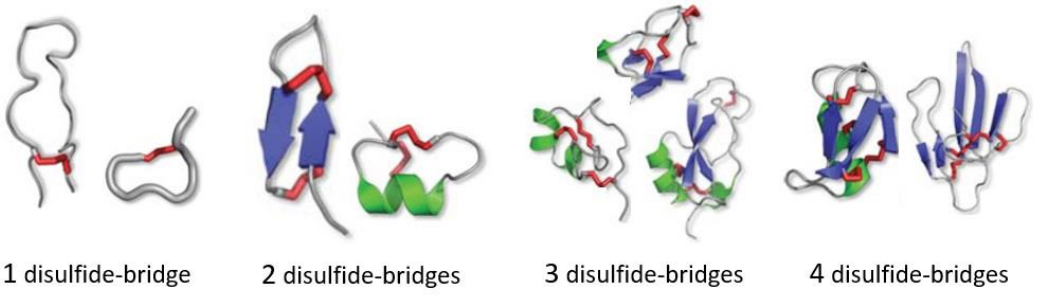
2.2. Venom peptides

The most important bioactive molecules present in animal venoms are small peptides, with no more than 120 amino acids, and reticulated by at least one disulfide bridge. Venom peptides have extremely important properties, such as stability conferred by conserved disulfide-rich scaffolds, and specificity and potency for a range of molecular targets. In addition, to escape the defence mechanism of the prey, reticulated peptides have been evolutionary fine-tuned to display low immunogenicity. It is widely recognized that the structural and functional diversity of venom toxins made them a valuable source of natural products for drug discovery.

2.2.1. Disulfide-rich venom peptides

Previous studies have demonstrated that the majority of venom peptides contain a rather high number of cysteine residues in their structure that oxidize to stabilise the conformation of these peptides through the formation of disulfide bonds. Venom peptides are highly stable to survive chemical degradation in solution at ambient temperature and enzymatic degradation by proteases present in the venom itself or in the tissues of prey species. This stability is often achieved naturally with post-translational modifications (PTMs). Disulfide bonds allow the peptide to fold into a highly stabilized and bioactive structure (Lewis & Garcia, 2003). Examples of others PTMs present in venom peptides are C-terminal amidation, N-terminal pyroglutamate and L-to-D isomerization of key amino acids residues (Buczek, Bulaj, & Olivera, 2005; King, 2011; Lewis & Garcia, 2003). However, the most conserved structural feature of venom peptides is the presence of a high number of intra-chain disulfide bonds, in relation to their backbone length (Figure 2.4). Thus, the formation of disulfide bonds contributes to the stabilization of their tertiary structure and confers rigidity to the molecules, as well as resistance to denaturation.

Figure 2.4| Examples of disulfide bridges in venom peptides. Disulfide bridges between two cysteine residues are shown as red tubes.



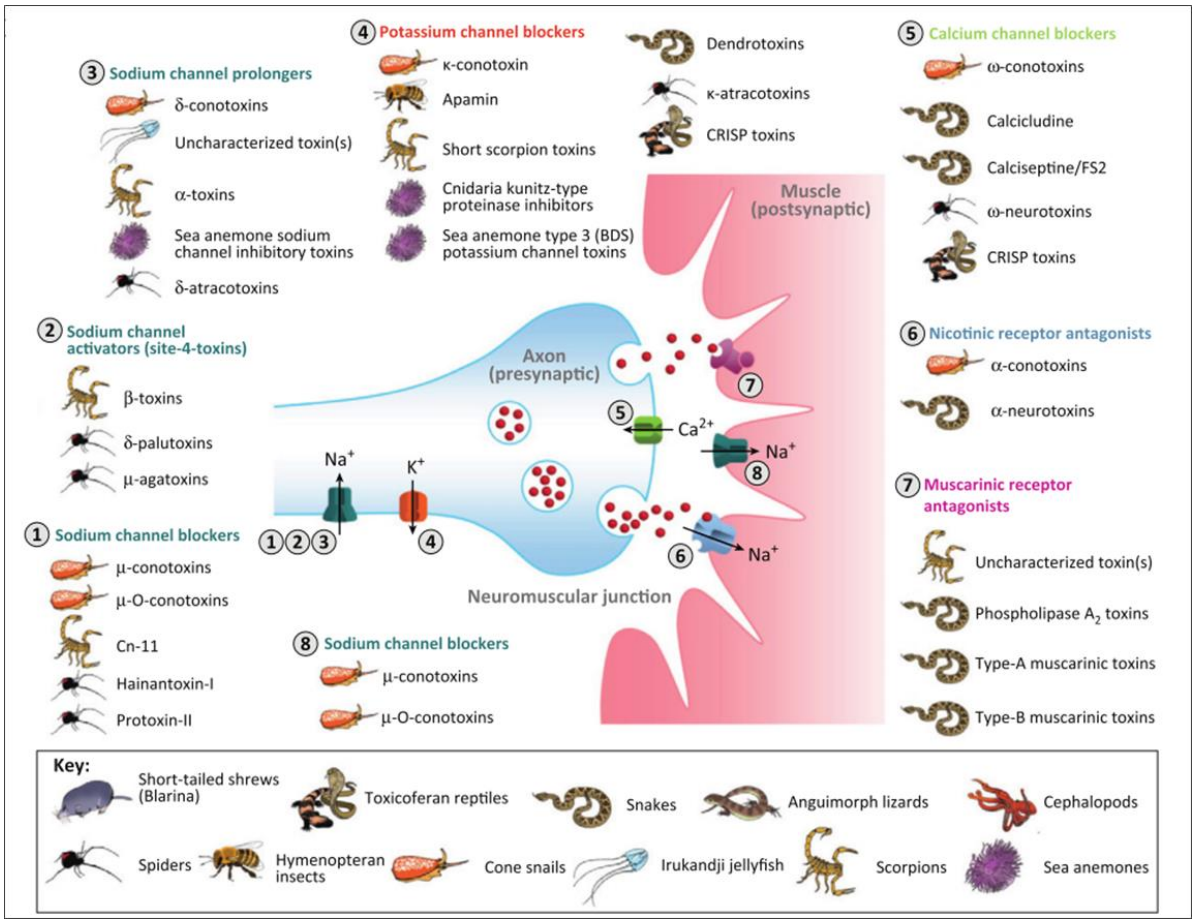
The final disulfide bridge pattern of venom peptides increases the efficacy of receptors ligation, through strict positioning of peptide residues that are key to interaction with membrane receptors (e.g. ion channels) (Mouhat, Jouirou, Mosbah, De Waard, & Sabatier, 2004). In addition, the number of cysteine residues and the cysteine-rich framework, in which all cysteine residues are involved in intra-chain disulfide bonds, are important features to aggregate venom peptides in classes. Based on this principle, the peptides isolated from the venom of cone snails, for instance, are separated in two main groups. The first represents peptides with more than two cysteine residues while the second includes peptides with a single or no cysteine group. Currently, there are twelve superfamilies of conus peptides and each superfamily contains several subfamilies that are characterized by their pharmacological activity (Kaas, Westermann, Halai, Wang, & Craik, 2008). There are also families of venom peptides that contain unique disulfide frameworks that are only found in specific animal taxa. For example, scorpion toxins are dominated by cysteine stabilized α/β and their core framework is the C-C-

C-C-C-C pattern. The chlorotoxin is a 36-residue CSα/β, isolated from the venom of Egyptian scorpion *Leiurus quinquestriatus quinquestriatus* and contains eight cysteine residues distributed in a C-C-C-CC-C-C-C pattern (Lavergne, Alewood, Mobli, & King, 2015). These structural and functional properties of venom peptides contribute to increase the diversity of venoms and to fine-tune the specificity for molecular targets.

2.2.2. Targets and function of venom peptides

Venom peptides have high affinity and selectivity for a diverse number of biological targets, especially membrane proteins such as ion channels, receptors (e.g. G-protein coupled receptors – GPCRs) (Näreoja & Näsman, 2012) and transporters. Many venom peptides target the nervous system by acting at various molecular sites such as central or peripheral neurons, axons, the synapse or the neuromuscular junction. Ion channels targeted by venom peptides display extensive structural diversity, including voltage-gated ion channels (VGCs), potassium, sodium and calcium channels, and the ligand-gated ion channels (e.g. nicotinic acetylcholine receptors). In addition, a set of venom peptides have evolved to target the N-methyl-D-aspartate receptor (NMDA receptor), chloride channels or noradrenaline transporter (NET) (Dutertre & Lewis, 2010; Escoubas *et al*., 2008; Lewis & Garcia, 2003). Ion channels are large protein complexes located in biological membranes and play a key function in the generation, shaping and transduction of electrical signals of neurons and other excitable cells. (Catterall, 1995; Terlau & Olivera, 2004). Recent studies revealed that venom peptides act in the vertebrate nervous system pre- and pos-synaptically and these biomolecules have played a vital role in our understanding of pain and neuronal signalling (Trim & Trim, 2013). The venom peptides that participate in disruption of neurotransmission by ligation to membrane receptors, both pre-synaptically (sodium, potassium and calcium channels) and post-synaptically (muscarinic and nicotinic receptors) are defined as neurotoxins. Many venomous arthropods such as bees, wasps and ants, spiders, centipedes and scorpions possess venom peptides that act on ion channels. In addition, peptides from cnidarians and cone snails block voltage-gated sodium and potassium channels of the pain pathway (Figure 2.5) (Casewell *et al*., 2013). These examples demonstrate that neurotoxins evolved different functions in the neurological pathway such as sodium/potassium/calcium channel blockers, sodium channel activators and prolongers, nicotinic receptor antagonists or muscarinic receptor antagonists (Trim & Trim, 2013).

Figure 2.5| Neuronal binding sites of neurotoxins identified in venoms of different venomous species.



Each number represents a different physiological target that is addressed by different neurotoxins presents in venoms of animals. Figure adapted from Casewell *et al.* (2013).

2.3. Synthetic strategies for venom peptides

Recent technological advances have facilitated the screening and structural characterization of venoms and venom peptides. However, the most significant bottleneck in venom research is the rapid production of non-limited quantities of venom peptides for complete structural and functional characterization. Most venomous animals do not produce large quantities of venoms and generally this native material is insufficient for bioactivity assays. This, obviously, does not apply for snakes that typically provide larger quantities of venoms than any other venomous animal (e.g. centipedes, scorpions, spiders or other smaller animals) (King, 2011; Vetter *et al*., 2011). The alternative is to produce in laboratory these venom peptides by one of two ways: using the chemical method of solid phase peptide synthesis (SPPS), or by recombinant expression of the peptide in a bacterial system or other type of heterologous host. There are advantages and disadvantages to SPPS and recombinant production of venom peptides, and the choice depends on properties such as the peptide length (SPPS is more appropriate for smaller peptides) and the presence and absence of post-translational modifications (recombinant production in bacterial hosts is not adequate for some PTMs).

2.3.1. Solid Phase Peptide Synthesis (SPPS) of venom peptides

SPPS is a chemical synthesis method that allows to synthesise peptides on a solid support, or resin, where the peptide sequence is assembled by coupling each amino acid one by one. The first amino acid is attached at its C-terminus to a resin via a linker and subsequent elongation of the peptide occurs with immobilization on the support until cleavage. This synthesis process involves repeated cycles of amino acid coupling into the growing polypeptides and deprotection. Upon conclusion of peptide synthesis, the synthetic peptide is cleaved from the resin and the linear peptide requires the subsequent step of oxidative folding in which the formation of disulfide bonds is promoted by air oxidation. Other PTMs can be introduced after SPPS are concluded. This method is the predominant approach employed for the production of short peptides and when non-natural modifications or exotic PTMs are required. This makes SPPS particularly suitable for production of cone snails peptides (< 40 residues), where 77% of all structurally characterized peptides from conus venoms contain at least one PTM other than disulfide bonds (Kaas *et al*., 2008; Vetter *et al*., 2011).

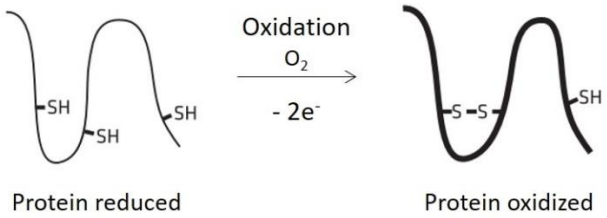
2.3.2. Recombinant venom peptides expression in *Escherichia coli*

Venom peptides can also be produced by overexpression in a prokaryotic or eukaryotic host systems. The Gram-negative bacterium *Escherichia coli* is the most widely used recombinant expression system and generally the first choice for recombinant venom peptide production. However, other eukaryotic hosts, such as yeast or insect cells, have been used for the same purpose (Vetter *et al*., 2011). Heterologous proteins are generally expressed in the cytoplasm of *E. coli* allowing to obtain high protein yields. However, the bacterial cytoplasm is predominantly a reducing environment, which is not appropriate for the folding of disulfide-rich peptides. Several studies suggested that bacterial hosts lead to low yields of correctly folded protein, and consequently, the majority of the recombinant peptides are sequestered into aggregates known as inclusion bodies, in the reduced and unfolded forms (de Marco, 2009; Klint *et al*., 2013).

The process of disulfide bond formation between cysteine residues, which are abundant in venom peptides, requires specific conditions. Cysteine (Cys) is a very important amino acid since it allows the formation of internal disulfide bonds, which are very important to stabilize the biologically active conformation of peptides. Cys residues have diverse functions in proteins: i) they stabilize protein folding through disulfide bonds; ii) they are part of active sites of enzymes; iii) they belong to regulatory sites of enzymes and proteins; and iv) they bind metals (Salinas, Pellizza, Margenat, Fló, & Fernández, 2011). The formation of a disulfide bond between two Cys residues is a redox reaction, in which the two thiol groups are oxidized to a disulfide in a reaction that releases two electrons (Figure 2.6). Within the cell, there are different compartments that possess distinct redox potentials and biochemical pathways that keep Cys residues either in a reduced or in an oxidized state. In eukaryotic cytosol and bacterial

cytoplasm, the environment is reduced, where specific pathways maintain Cys residues in a reduced state and the oxidation of thiol groups is disfavoured. The redox homeostasis that maintain Cys residues in a reduced form in the cytoplasm of bacterial cells is controlled by the Glutathione/Thioredoxin system. In contrast, the eukaryotic endoplasmic reticulum (ER) and the bacterial periplasm are oxidative environments, which include enzymes that participate in the formation of disulfide bonds. Therefore, the ER and the bacterial periplasm provide adequate environments for disulfide-bond formation.

Figure 2.6| Formation of disulfide bonds between two Cys residues.



Taken the above, the low levels of expression of recombinant rich-disulfide peptides in the *E. coli* cytoplasm may result from the unfavourable and reducing environment of this cellular compartment. Thus, expression of these peptides in bacterial cytoplasm promotes that all cytoplasmic cysteine residues are kept in the reduced state (Salinas *et al*., 2011) and, consequently, the formation of disulfide bond by oxidation is disfavoured. Based on this principle, the choice of a favourable environment for disulfide bond formation is an important consideration in experimental design, in order to make *E. coli* strains more suitable for expression of disulfide-rich peptides. In past years, many approaches have been attempted to promote the formation of disulfide bonds and the native folding of disulfide-rich peptides. Presently two main approaches have been developed: the first is based on adjustments in cytoplasmic compartment to make it less reducing, and the alternative method comprehends the exportation of the recombinant peptide to the oxidized *E. coli* periplasm (Klint *et al*., 2013; Vetter *et al*., 2011). To facilitate folding and to avoid the formation of inclusion bodies, cysteine-rich peptides are often expressed in *E. coli* strains with defective glutathione and thioredoxin reductases. Therefore, the ablation of this redox pathway leads to an oxidizing environment in bacterial cytoplasm and consequently results in the production of higher levels of folded peptides. The cytoplasmic accumulation of correctly folded disulfide-rich peptides can also be improved by the co-expression of the protein disulfide isomerases, such as protein disulfide isomerase (PDI) and disulfide-bond isomerase (DsbC). The last isomerase, DsbC, enhances the fidelity of disulfide bond formation and helps protein folding due to its general chaperone activity (Salinas *et al*., 2011; Klint *et al*., 2013; Nozach *et al*., 2013b). To overcome this issue, engineered strains with an oxidized cytoplasm have been developed, like *E. coli* SHuffle® strain. These *E. coli* strains contain deletions of both glutathione and thioredoxin redutase genes (*gor*$^-$, *trxB*$^-$) and express the disulfide-bond isomerase DsbC within the cytoplasm

(Berkmen *et al*., 2012). The alternative approach, or a more effective approach for Klint *et al*. (2013), is to direct the recombinant peptide to the *E. coli* oxidizing periplasm, where the endogenous protein machinery for disulfide bond formation is located (Vetter *et al*., 2011; Klint *et al*., 2013). This endogenous machinery includes four thiol-disulfide oxidoreductases known – DsbA, DsbB, DsbC and DsbD; the DsbA is the most oxidizing protein and quickly reacts with unfolded proteins as they enter in periplasm (Choi & Lee, 2004; de Marco, 2009; Salinas *et al*., 2011). The exportation of proteins to *E. coli* periplasm involves the introduction of a periplasmic export sequence (or signal peptide) at the N-terminus of the desired protein, such as the *MalE* signal sequence. Other signal peptides have been successfully and widely used for this purpose, like M13 pIII or signal peptides from periplasmic proteins, such as DsbA (Salinas *et al*., 2011). In summary, to export proteins to *E. coli* periplasm can be an intuitive strategy. However, the secretion of proteins to the periplasm often leads to low protein production, probably because of the limited periplasmic volume combined with an insufficient capacity of the translocation machinery. Since levels of protein expression are important in recombinant production, many researchers have chosen to follow strategies based on production of rich-disulfide peptides in a modified *E. coli* cytoplasm, when higher protein yields are absolutely required (Nozach *et al*., 2013).

Eukaryotic expression systems have also been used for the production of disulfide-rich peptides. However, mammalian and insect expression systems are costly, time consuming and produce low protein yields (Escoubas, Bernard, Lambeau, Lazdunski, & Darbon, 2003). In contrast, yeast has been revealed to be an excellent system for expression of disulfide-rich peptides presenting several advantages, such as high yields, low cost and the ability to incorporate PTMs (Wu *et al*., 2002). Moreover, yeast cells possess the ability to secrete proteins directly into the yeast growth medium, a process that simplifies subsequent protein purification steps. *Saccharomyces cerevisiae* and *Pichia pastoris* are the most used yeast, being the later an excellent system for the production of proteins with five or more disulfide bonds (Macauley-Patrick, Fazenda, McNeil, & Harvey, 2005). To resume, there are several strategies for the recombinant production of disulfide-rich peptides, being the periplasmic expression in *E. coli* the best choice for many authors (Klint *et al*., 2013; O'Reilly, Cole, Lopes, Lampert, & Wallace, 2014; Vetter *et al*., 2011), while the extracellular secretion in *Pichia pastoris* a backup of *E. coli* expression system. SPPS is usually the choice for the production of smaller peptides. This thesis reports the development of a system to improve the production of disulfide-rich peptides in *E. coli*, using an optimized codon usage and an efficient expression vector that carries the periplasmic DsbC to ensure high recombinant protein levels with proper fold (see Chapter 5).

2.4. Novel approaches for production of venom peptides in post-genomics era

DNA sequencing technologies have undergone remarkable improvements since the establishment of The Human Genome project (HGP) initiated in 1990 (http://www.genome.gov/10001772). Informatics solutions have helped advancing specific fields of research that started dealing with huge amounts of information, in particular those related with biology, genetics and genomic, by developing innovative methods to manage big amounts of data (Staden, 1979; Weber & Myers, 1997). In addition, the HGP fostered the development of novel sequencing technologies that improved the efficiency of DNA sequencing. This next generation sequencing (NGS) technologies were based on high-throughput platforms allowing the extensive characterization of genomes and metagenomes and rapidly became the routine DNA sequencing technology to address big projects. Several genome sequencing projects are presently ongoing, including the sequencing of bacteria, virus, plants and animals. Genomes of a few venomous animals were also sequenced to understand how these animals have evolved to play a diversity of physiological functions, such as the production and delivery of their venoms, and the envenomation process. The genome sequence of *Mesobuthus martensii*, a venomous scorpion from Asian countries, is already available, providing a complete analysis of 32,016 protein-coding genes. The complete sequencing of *M. martensii* genome allowed to explore the genetic contributions underlying the long-term survival and adaptive model of scorpions, which contributed to understand the nocturnal behaviour, feeding and prey capture of these animals (Cao *et al*., 2013). Other examples of sequenced genomes of venomous animals are from *Apis mellifera* (honey bee) and *Solenopsis invicta* (fire ant), a bee and an ant from the venomous insects group, whose genomes were sequenced in 2006 and 2011, respectively (Weinstock *et al*., 2006; Wurm *et al*., 2011).

Advances in sequencing technologies have generated an enormous amount of genetic information that is highly relevant for all biological sciences. As a consequence, not only new tools to manage large amounts of data need to be developed but also it is particularly important to decipher unknown information encoded by genomes and metagenomes. Thus, the huge challenge of the post-genomics era is to assign a biological function to all genes/proteins encoded by genomes (Eisenberg, Marcotte, Xenarios, & Yeates, 2000). Target genes that encode unknown proteins can be discovered from genomic databases using bioinformatics tools, then subsequently amplified by the Polymerase Chain Reaction (PCR) with specific primers or synthetically produced, and cloned into appropriate expression systems. The PCR from a genomic DNA template is a simple and widely used procedure but has several limitations when applied to eukaryotic genomes. In many cases, the chemical synthesis of gene sequences may be the best choice because template DNAs are often not readily available due to limiting amounts or the difficult access to biological material from

extremophiles organisms (e.g. psychrophiles, thermophiles, acidophiles organisms) or small animals (e.g. small venomous animals as bees, ants, small spiders), or when the expression of the natural gene in heterologous systems like *E. coli* may not be ideal. Other situations where the chemical synthesis is presently the best strategy for DNA isolation are when the codon optimization for specific expression system is required or when proteins need to be engineered for higher stability and/or more potent activities. The approach to obtain the target genes has been revolutionized with the introduction of Synthetic Biology in strategies underlying protein research. Scientists gained the ability to write genetic information encoding any protein that must be expressed and analysed, using protocols to synthesise *de novo* DNA constructs with any size or sequence. The technology to write DNA is termed 'gene synthesis' and appears as a paradigm change in the field of recombinant protein production and biology. This approach has proven to be the best way to obtain target genes that encode peptides identified in venoms of small animals, since the biological material of these animals is reduced and of difficult access (Venomics, 2012). Moreover, the use of this approach allows the optimization of DNA sequences which encode eukaryotic proteins for an efficient expression in host systems that can be, for example, prokaryotic systems like *E. coli*.

2.4.1. Synthetic biology

Synthetic biology is a young discipline that combines elements of biological sciences, engineering and computational modelling, using tools and concepts that help the synthetic biologists to design new biological systems that mirror natural biology. The term 'synthetic biology' appeared for the first time in 1980 to describe bacteria that had been genetically engineered using recombinant DNA technology (Hobom, 1980). Benner and Sismour (2005) describe synthetic biology as an assembly of components that are not natural (synthetic) to generate chemical systems that support Darwinian evolution (biological) (Benner & Sismour, 2005; Way, Collins, Keasling, & Silver, 2014). Other scientists adopt the definition of synthetic as artificial, unnatural, or not occurring in nature (Tian, Ma, & Saaem, 2009). This field has rapidly emerged over the last years by developing new biological components. Thus, it is possible to explore biological functions, expression of target genes or analyse the action of engineered proteins in biochemical pathways with the goal of understanding functional mechanisms in living systems (Benner & Sismour, 2005). Synthetic biology has created living organisms with non-natural components or based on alternative genetic codes, as well as minimal organisms based on cells which contain the minimum number of genes essential for life, thus representing blank cells for inputting new functions (Polizzi, 2013). Other example of these projects are the directed evolution studies that use engineering to modify proteins, providing pools of polypeptides with partly randomized sequences, which are then screened to identify the desired variation (Bershtein & Tawfik, 2008). Recently, new tools for genome engineering have been developed by synthetic biologists, such as methods based on zinc-

finger nucleases, transcription activator-like effector nucleases (TALE) and clustered regularly-interspaced short palindromic repeats (CRISPR), which can be used to generate double-strand breaks at specific sites in the genome (Gaj, Gersbach, & Barbas, 2013).

2.4.1.1. Gene Synthesis: designing genes for successful protein expression

The novel capacities created by synthetic biology to generate novel molecules have improved our capacity to recombinantly produce functional proteins in heterologous hosts. Unfortunately, proteins are often difficult to express outside their original context. In many cases, the target protein is not expressed or is expressed only at very low levels. Much work has been done to improve the expression of cloned genes (genes cloned from cDNA libraries or by PCR), including optimization of host growth conditions and the development of new host strains. However, these approaches have not proved to be a unique solution of the problem underlying the lower expression levels. There is considerable evidence suggesting that the DNA sequence of gene encoding the target protein can also have a dramatic influence in expression levels; very often the DNA sequence encoding the recombinant protein to be express in one organism is quite different from the sequence of endogenous genes (Gustafsson, Govindarajan, & Minshull, 2004). These proteins might contain codons that are rarely used in the selected host, or contain expression-limiting regulatory elements within their coding sequence. The use of a designed gene (synthetic gene) which codes for the target protein can enhance the gene expression levels in the host, for example by matching the codon usage with that of the host in which the gene is expressed or by removing specific sequences that are unfavourable for protein expression. Several studies have revealed that the redesign of the entire gene sequence has direct implications in the levels of expression of heterologous proteins (Gustafsson *et al*., 2004; Welch, Villalobos, Gustafsson, & Minshull, 2009). Thus, *de novo* gene synthesis emerges as a new tool to redesign genes and to create novel elements that are not existing in nature, like unnatural genes. This approach corresponds to a new application of genetic engineering providing a powerful tool for producing and modifying genes and to explore their structure, expression and function. Synthetic genes can be used to express proteins of interest in a host cell, making it possible to produce high levels of heterologous proteins. Designing an optimal gene requires a deep understanding of the interaction of the gene sequence with the expression host. There is no simple formula to guarantee success. Nevertheless, several steps can be taken to greatly increase the probability that a desired DNA sequence will result in expression of the encoded protein. In general, a robust gene design method involves the adaptation of codon usage of the synthetic genes to the genetic code of the host organism. This process is termed “codon optimization” and it is based on the degeneracy of the genetic code.

2.4.1.2. Sequence parameters affecting protein expression

Codon optimization is a technique recently used by many scientists to improve recombinant protein expression in living organisms by increasing the translational efficiency of the gene of interest. Several synthetic gene design strategies have been developed to mimic natural gene characteristics that are relevant for improved expression. These strategies combine the genetic information of the protein to be expressed with codon usage of the host system, and other important factors underlying efficient expression of the desired protein. Codon usage has been identified as the single most important factor in prokaryotic gene expression (Lithwick & Margalit, 2003). However, other factors were also shown to play some role in efficient gene design. These factors include presence of messenger RNA (mRNA) secondary structures around translation initiation region and the strength of the interaction between the ribosome and the Shine-Dalgarno (SD) sequence of mRNAs, which can compromise the efficiency of the translation process. The incorporation or removal of recognition sites for restriction enzymes that enable DNA manipulations techniques, such as subcloning into expression vectors, is another factor to be considerate in gene design process. Repeated sequences, potential polyadenylation sites, cryptic splice sites, nuclease cleavage sites, stop codons, and guanine-cytosine (GC) content are important factors that are also known to affect gene expression (Welch, Villalobos, Gustafsson, & Minshull, 2011). Some of these factors can be incorporated and/or eliminated in synthetic genes using computational tools.

2.4.1.2.1. Codon bias

The standard genetic code uses 61 nucleotide triplets (codons) to encode 20 amino acids and three codons to terminate the translation. Each amino acid is therefore encoded by between one (methionine and tryptophan) to six (arginine, leucine and serine) synonymous codons. This degeneracy of the genetic code enables many alternative nucleic acid sequences to encode the same protein. For example, a 300 amino acid protein of average amino acid composition could be encoded by more than 10^{100} different gene sequences (Welch, Villalobos, *et al*., 2009). The frequencies by which different codons are used vary significantly between different organisms, meaning that each organism have evolved to work with a particular set of codons (termed codon usage). Significant differences between codon usages of different organisms is often termed codon bias. In addition, codon usage of an organism is correlated with the availability of transfer RNA (tRNA) molecules within the cell. Thus, rare codons for a given amino acid are usually correlated with a reduced intracellular level of tRNA molecules and these codons should be avoided during the codon optimization process. Other important concept is the codon adaptation index (CAI) that was originally proposed by Sharp and Li (Sharp & Li, 1987) based on the premise that each amino acid has a “best or preferred” codon for a particular organism. The CAI is derived from a reference set of highly expressed genes used to score the preference of an organism for specific codon. Ratios between the

frequency of each codon and the preferred synonymous codon frequency can be calculated as a CAI score for a given transcript. CAI values can vary between 0 and 1, with CAI values approximating 1 potentially correlated with high expression levels. Although the CAI of a gene has often been cited as a predictor of the expression level of a protein, there is no demonstrated causality. Studies using the *E. coli* and *Saccharomyces cerevisiae* host systems, did not identify correlations between CAI values and protein yield per mRNA transcript suggesting that CAI may not always constitute the best tool to measure the translational efficiency (Welch, Villalobos, *et al*., 2009).

There are different approaches to design a gene based on CAI score. The most traditional approach is related with a high value of CAI in which the most frequent codon corresponds to the highest translation efficiency in heterologous expression. Based on this principle several algorithms were developed to optimize DNA sequences relying on the maximization of CAI values, meaning that the most frequent codon for an amino acid is always the selected codon for that amino acid. In opposition to this perspective, Mark Welch *et al*. (2009) suggest three reasons explaining why the "best or preferred" codon approach to gene design may inhibit protein expression: 1) overuse of the "best" codon for a given amino acid could result in high usage of only a subset of the tRNA pool, possibly exhausting their availability and could result in translational errors; 2) no flexibility in codon usage could make it impossible to avoid repetitive elements and secondary structure of mRNA molecules, possibly inhibiting ribosome processing; and 3) if codon usage is rigidly fixed, inclusion or exclusion of restriction sites relevant to gene synthesis may be impossible (Welch, Govindarajan, *et al*., 2009). These authors developed a codon-usage model based on this alternative approach that is provided free of charge to clients requiring gene synthesis to the company DNA2.0 (California). Other studies have shown the same evidence, that genes designed using preferred codons are not correlated with high protein expression and these high levels can be also related with weak mRNA structures (Kudla, Murray, Tollervey, & Plotkin, 2009). Allert *et al*. (2010) designed 285 synthetic genes in order to analyse the influence of CAI values, adenine-thymine (AT) content and mRNA structures in expression levels in *E. coli*. The data revealed that increasing AT content at the extremes of gene sequence, particularly at the 5' end, improves the expression levels of targeted protein (Allert, Cox, & Hellinga, 2010).

2.4.1.2.2. Translation and mRNA structure

The general principle of protein expression refers to the way in which information of a gene is used in the synthesis of a functional protein. In others words, DNA is transcribed to mRNA, which is translated to protein. The efficiency of translation, and the resulting level of protein production is determined by both translation initiation and elongation rates. Translation elongation rate controls the speed of the translation process through ribosome density profiles (Quax, Claassens, Söll, & van der Oost, 2015). In prokaryotic systems, the initiation of

translation occurs in ribosome binding site (RBS) that is located between 5 and 15 bases upstream of the open reading frame (ORF) AUG start codon. This short region of mRNA, called the Shine-Dalgarno (SD) sequence, drives the ribosome to the initiation codon by ligation with the anti-SD sequence localized in 16S rRNA. The binding strength between the SD sequence in mRNA and anti-SD sequence in ribosome regulates the efficiency of the translation initiation. Thus, the affinity of the RBS for the ribosome is a critical factor for the initiation of the translation process. Alterations in RBS sequence can change expression levels over more than three orders of magnitude (Welch *et al*., 2011). Several authors have demonstrated that mRNA structures that obstruct the RBS region and/or the start codon in genes expressed in prokaryotes can significantly reduce protein expression, probably by interfering with ribosomal binding and translational initiation. Thus, it is critical to minimize the formation of mRNA secondary structures in gene design strategies, namely around the regulatory initiation sequences and start codon. In addition, there is a considerable degree of evidence suggesting that the initial 15-25 codons of the open reading frame deserve special consideration in gene optimization (Allert *et al*., 2010; Welch *et al*., 2011; Welch, Villalobos, *et al*., 2009).

2.4.1.2.3. Algorithms for codon optimization

Over the last years significant advances in several innovative computational tools allowed creating novel possibilities to design synthetic genes for optimized protein expression. Different algorithms have been developed to best adapt a coding gene to the codon usage of the host organism. The most common optimization strategy is based on the removal of rare codons and in the maximization of the most frequent codons. Other codon optimization strategies that are not based on this principle are also implemented in computational tools. Presently, there are 13 software packages available to design optimized DNA sequences for protein expression in host system (Table 2.1): DNAWorks, Jcat, Synthetic Gene Developer, GeneDesign, Gene Designer 2.0, OPTIMIZER, Visual Gene Developer, Eugene, Codon Optimization Online (COOL), D-Tailor, CodonOpt, GeneGenie and ATGme. These computational tools vary in the types of design criteria they support and in the codon optimization techniques. Some of these bioinformatics tools can also provide oligonucleotides design, such as DNAWorks, Synthetic Gene Developer, GeneDesign, OPTIMIZER and GeneGenie. The design criteria may include: 1) codon usage tables for one or more host systems, 2) minimization of the secondary structures of mRNA, 3) avoiding rare codons, 4) insertion/removal of restriction sites for restriction enzymes, 5) avoiding repeated sequences, potential polyadenylation sites and cryptic splice sites, and 6) adjustments of GC content. For example, DNAWorks is a web-based application that provides codon optimization and oligo generation. It was originally created to improve the process of oligonucleotide design for synthetic gene construction. Moreover, it displays tools to adjust codon utilization, although presenting some limitations. ATGme is a more recent application for gene optimization and was developed by Daniel *et al*.

(2015) as a user-friendly open-source web-based application. This application allows the identification of rare codons in the DNA sequence to be optimized and offers three different methods for gene optimization. However, ATGme provides a highly simplified codon optimization method, since there is no tool for back-translation, only provides the user the possibility to look at the sequence and to change each codon one by one. The option to address any codon by itself can only be found in two available programs: CodonOpt and ATGme. CodonOptimization from IDT® Technologies provides the possibility of designing multiple genes at the same time, which makes it the unique available tool compatible with high-throughput gene synthesis methods.

Table 2.1| Gene design tools.

Gene design tool	Web URL	Reference
DNAWorks	http://helixweb.nih.gov/dnaworks/	Hoover and Lubkowski (2002)
Jcat	http://www.jcat.de/	Grote *et al*. (2005)
Synthetic Gene Developer	http://userpages.umbc.edu/~wug1/codon/sgd/	Wu *et al.* (2005)
GeneDesign	http://genedesign.org/	Richardson *et al*. (2006)
Gene Designer 2.0	http://www.dna20.com/resources/genedesigner	Villalobos *et al.* (2006)
OPTIMIZER	http://genomes.urv.es/OPTIMIZER	Puígbo *et al.* (2007)
Visual Gene Developer	http://www.visualgenedeveloper.net/	Jungo & MacDonald (2011)
Eugene	http://bioinformatics.ua.pt/eugene	Gaspar *et al.* (2012)
Codon Optimization Online (COOL)	http://bioinfo.bti.a-star.edu.sg/COOL/	Gaspar *et al.* (2013)
D-Tailor	http://sourceforge.net/projects/dtailor	Chin *et al.* (2014)
CodonOpt	https://eu.idtdna.com/CodonOpt	IDT® Technologies
GeneGenie	https://www.gene-genie.org	Swainston *et al.* (204)
ATGme	http://atgme.org	Daniel *et al.* (2015)

According to the different issues that were discussed above it is clear that the process for designing an optimal gene that ensures high expression levels is still not well defined. Evidences that recoding a gene using preferred host codons will not maximize protein expression were demonstrated in different studies. Moreover, the removal of mRNA structures can be important in codon optimization process, as well as other factors such as the elimination of rare codons or repetitive sequences. However, these results are only indications of the best strategy to design a gene encoding the desired protein; there is still a long path to reach the formula for guaranteed success underlying protein expression. This thesis reports the development of a robust algorithm for gene design using codon usage tables for *E. coli* (see Chapters 3 and 5). The creation of this codon optimization algorithm combines factors that influence the protein expression with a new form to select codons from the *E. coli* usage table. This algorithm intends to revolutionize the capacity to design multiple genes, being compatible with HTP methodologies. It was successfully used to design hundreds of eukaryotic genes

encoding venom peptides that were synthetically produced by PCR-based methods (see Chapter 6).

2.4.1.3. Methods for gene synthesis

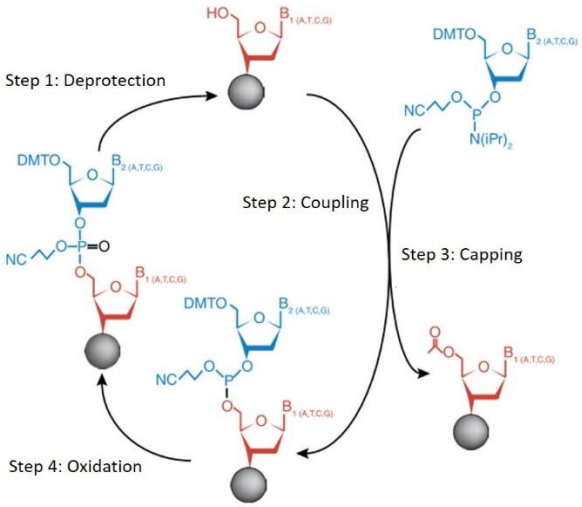
Once a gene sequence has been designed according to the parameters described above, the synthetic gene can be produced using different approaches. Synthetic biology is a powerful resource to develop protein engineering projects that are both structurally and functionally relevant. In general, gene synthesis is a chemical method that uses synthetic oligonucleotides and different methods of assembling to obtain fragments of double-stranded DNA that are usually cloned into a plasmid vector. The first example of a synthetically produced gene was presented in 1970 by Khorana and colleagues (Agarwal *et al*., 1970). In those pioneering studies, a gene of 77 bp encoding yeast alanine transfer RNA was produced using short oligonucleotides obtained by organic chemistry methods. Advances in the technology of oligonucleotide synthesis and the subsequent decrease in prices promoted the development of different gene synthesis methods, using T4 DNA ligase (Edge *et al*., 1981), heat stable ligases (Barany & Gelfand, 1991) and the ligase chain reaction (LCR) (Young & Dong, 2004). The invention of the polymerase chain reaction (PCR) in 1985 (Saiki *et al*., 1985) enabled the development of new and more efficient methods to produce synthetic genes based on PCR.

2.4.1.3.1. Oligonucleotides synthesis

The core of all gene synthesis methodologies relies on the chemical synthesis of oligonucleotides, the building blocks of synthetic genes that are subjected to enzymatic assembly. The most frequently used approach for the synthesis of oligonucleotides is the four-step phosphoramidite synthesis method, developed in the early 1980s (Caruthers *et al*., 1983, 1987). Oligonucleotide synthesis is a cyclical process that elongates a chain of nucleotides from 3' to 5' direction by coupling acid-activated deoxynucleoside phosphoramidites to initial deoxynucleoside attached to a solid support through a 3'-hidroxyl group. Most commercial synthesisers use controlled pore glass (CPG) or polystyrene (PS) as the solid support which is packed into a flow-through column. The addition of each nucleotide monomer to the growing oligonucleotide chain is performed in four steps (Figure 2.7): (1) deprotection: a dimethoxytrityl (DMT) ether group is removed by washing with a weak acid from the 5'-hydroxil (OH) end of the growing oligonucleotide chain, exposing the 5'-OH for the next coupling reaction; (2) coupling: the 5'-OH group generated from the deprotection step reacts with a tetrazole-activated monomer by simultaneous addition of the monomer and the activator solutions; (3) capping: any uncoupled 5'-OH groups are blocked by acylation to prevent later growth of an incorrect sequence; and (4) oxidation: the unstable phosphite triester internucleotide bonds are oxidized into more stable phosphotriester linkages. This cyclic process is repeated until the oligonucleotide chain is complete. Once the desired chain is produced it is cleaved from

the solid support and all the protecting groups are removed by treatment with a strong base such as ammonium hydroxide (Hughes, Miklos, & Ellington, 2011; Tian *et al*., 2009). An alternative two-step version of the phosphoramidite synthesis cycle was developed by Sierzchala and colleagues and simplified the above mentioned technology by eliminating several reactions (Sierzchala *et al*., 2003). This approach uses a carbonate group to substitute the DMT protecting group on the 5'-OH of each phosphoramidite. A peroxy anion is then used as a nucleophile in each synthesis cycle to simultaneously remove the 5'-carbonate protecting group and oxidise the internucleotide phosphite triester. Thus, this two-step method considerably simplified the synthesis procedure providing greater flexibility, more automation and less costs for the large scale synthesis of oligonucleotides (Sierzchala *et al*., 2003). Further improvements in throughput and reductions in costs drove to appearance of novel technologies for the synthesis of oligonucleotides, such as DNA synthesis on microarrays or microfluidic-devices (Tian *et al*., 2009). The chemical synthesis procedures described above are generally used for production of oligonucleotides shorter than 120-150 bases. However, longer oligomers of up to 300-600 bases have been synthesised at low yields and increased error rates. Accumulation of errors in longer nucleotides and the higher prices promoted the use of short oligonucleotides for the efficient production of synthetic genes.

Figure 2.7| Oligonucleotide synthesis using a four-step phosphoramidite synthesis cycle.



The synthetic process involves deprotection, coupling, capping and oxidation steps until the oligonucleotide chain is complete. Figure adapted from Kosuri *et al*. (2014).

2.4.1.3.2. Gene assembly methodologies

A variety of methodologies have been developed to assemble oligonucleotides into complete genes or large genomes. However, the most efficient and common technologies rely on a ligation-mediated assembly or a PCR-based assembly.

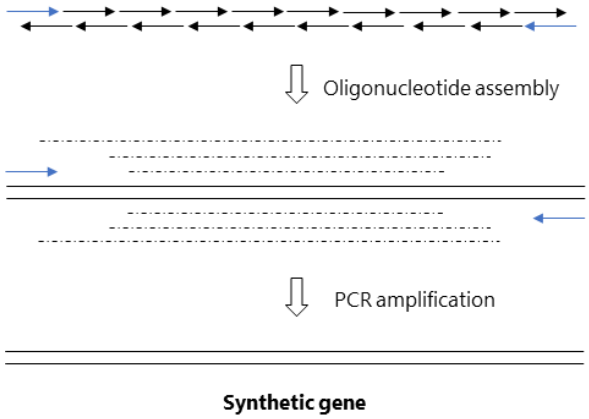
2.4.1.3.2.1. Ligation-mediated assembly

Assemblage of synthetic oligonucleotides using DNA ligases to construct genes is the earliest example of a gene synthesis procedure. Other ligation-mediated assembly methods have evolved from the original experiments, called Shotgun Ligation, and involve splitting of the gene product into multiple fragments composed of overlapping and phosphorylated oligonucleotides (Grundström *et al*., 1985; Eren & Swenson, 1989). The discovery of thermostable ligases has allowed the integration of the ligase chain reaction (LCR) in the synthesis of nucleic acids. Through LCR technology, a mixture of 5'-phosphorylated oligonucleotides, with overlap sequences that span both strands of a desired DNA duplex, is denatured and annealed together under different temperature cycles. The oligonucleotides are subsequently ligated by a thermostable DNA ligase to construct a double strand (ds) DNA. The gene product can be used as the template for additional ligation events until the desired gene fragment has been assembled. LCR is limited in its usage due to the necessity of phosphorylated oligonucleotides (more expensive) and it still requires the final synthesis assembly. Recently, novel technologies for the synthesis of multiple genes have emerged and those use the ligation-mediated assembly procedures enabling the rapid and cost-effective preparation of long DNA molecules. The Blue Heron technology, developed by Blue Heron Biotechnology company, is a gene synthesis platform that uses the ligation-mediated assembly method linked to a solid-phase support to produce synthetic genes for a broad market.

2.4.1.3.2.2. PCR-based assembly

The most commonly used gene synthesis technique is based on the PCR. This process is also termed templateless PCR due to the absence of a DNA template that is fundamental in conventional PCR reactions. PCR-assembly allows the construction of the desired DNA sequence from short oligonucleotides using thermal cycling reactions to obtain the fully assembled gene product. The first report of the polymerase chain assembly (PCA) to produce longer synthetic genes from overlapping short oligonucleotides was described by Stemmer *et al*. (1995). PCA is a version of the PCR process in which a thermostable DNA polymerase is used to stitch together oligonucleotides. Oligonucleotides with partial overlaps are pooled together and assembled in PCA reaction to form the gene of interest. The desired product is then amplified from the PCA reaction in a standard PCR using the outermost primers (Figure 2.8). Stemmer *et al*. (1995) used this method to produce a 1.1 kb β-lactamase encoding gene and a 2.7 kb plasmid using 40 bp overlapping oligonucleotides. Later, Withers-Martinez *et al*. (1999) used an optimized PCA method to synthesise a 2.1 kb gene from *Plasmodium falciparum*. The PCA method has become the most common approach to *de novo* gene synthesis although it does not work consistently for all genes and requires case-by-case optimization.

Figure 2.8| The PCR-based assembling method used by Stemmer *et al*. (1995) for production of a synthetic gene.

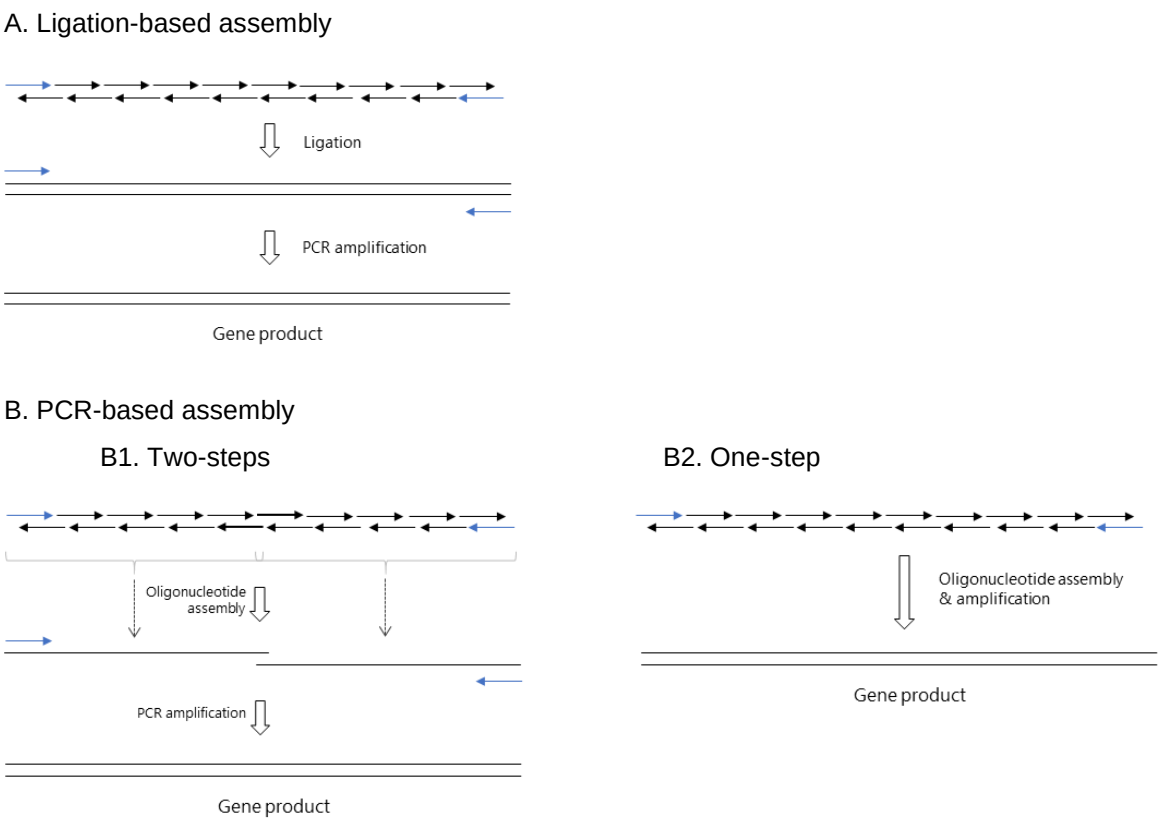


Overlapping oligonucleotides were designed without gaps to cover the entire gene sequence in both strands**.** The inner (internal) primers are represented by black arrows and the outer (external) primers are highlighted in blue. This method is initiated with the assembly of the oligonucleotides pool to produce a mixture of products with different lengths. The desired full-length product is then amplified by PCR using the outer primers.

Alternative methodologies for oligonucleotide design and PCR assembly were developed to increase the rates of successful PCA assembly, reduce error rates, and increase the throughput. These versions of PCA usually require multiple PCR reactions to build the desired gene in blocks prior to joining the blocks together by overlap extension. Therefore, long DNA sequences have been produced in two different steps (Figure 2.9, B1). Young and Dong (2004) developed a two-step PCR gene synthesis method that combines dual asymmetric PCR (DA-PCR) (Sandhu, Aleff, & Kline, 1992) and overlap extension PCR (OE-PCR) (Horton, Hunt, Ho, Pullen, & Pease, 1989). This method eliminated the requirement for optimization of reaction conditions and allowed a decrease in oligonucleotide cost since it used unpurified and unphosphorylated short oligonucleotides (< 25 bp) (Young & Dong, 2004). Xiong *et al*. (2004) developed another PCR-based two-step DNA synthesis (PTDS) method for gene synthesis of long DNA sequences that involves two steps: the first consists in synthesis of individual fragments, and the second step is the assembly of these individual fragments into the complete gene. The thermodynamically balanced inside-out (TBIO) (Gao, Yo, Keith, Ragan, & Harris, 2003), successive extension PCR and other improved PCR-based gene synthesis methods have been described and incorporate significant improvements to the earliest strategies of PCA (Hughes *et al*., 2011). As described above, most PCR assembly methods use two successive PCR steps to form the complete gene. However, Wu *et al*. (2006) have presented a simplified method that combines two steps in one (Figure 2.9, B2). With a single increase in outermost primers concentration, this author successfully synthesised three genes with different lengths (206, 777 and 936 bp) using an one-step PCR reaction for gene assembly (Wu *et al*., 2006). According to this technology, the parameters of the PCR reaction, in particular the choice of the DNA polymerase and the concentration of the assembled

oligonucleotides, are critical for successful gene synthesis and should be selected with precaution and accuracy. Moreover, this one-step approach proved to be an easier and cost-effectively method that is compatible with high-throughput protocols to efficiently synthesise multiple nucleic acids. PCA technique is also an appropriate approach to efficiently synthesise synthons that can be used as precursors to synthesise large constructs (Kodumal *et al*., 2004). For example, Gibson *et al*. (2008) synthesised a 583 kb *Mycoplasma genitalium* genome using *in vitro* recombination techniques to join overlapping "cassettes" with 5-7 kb, first assembled from synthetic oligonucleotides (Gibson *et al*., 2008).

Figure 2.9| Overview of most common enzymatic methods used for gene synthesis.



The most common enzymatic methods used for gene synthesis are Ligation-based assembly (A) and PCR-based assembly (B). A: ligation-mediated assembly usually involves two steps - ligation using DNA ligase and PCR amplification. B1: Two-step PCR-based assembly – the first step consists in the PCR assembly and extension of overlapped oligonucleotides while the second step is the amplification of full-length product using two outer oligonucleotides (blue). B2: One-step gene assembly reaction involves only the assembly of inner (black) with outer oligonucleotides and amplification for synthetic gene construction.

2.4.1.3.3. Why errors occur during gene synthesis

The main concern related with gene synthesis is the reduction of the number of errors identified in the synthetic DNA sequences. Sequence errors are usually incorporated during the gene synthesis process, either through the assembly of oligonucleotides containing errors or during the enzymatic assembling steps. Current oligonucleotide synthesis methods usually produce oligonucleotides that are prematurely terminated, or comprise internal insertions or deletions.

Deletions and insertions are the most common type of errors that can be introduced in synthetic oligonucleotides. Deletions can occur as a result of failures in capping or deprotection, with a frequency as high as 0.5% per position. While insertions are caused by unwanted DMT cleavage by tetrazole and can reach 0.4% per base (Ma, Saaem, & Tian, 2012). The chemical synthesis of a desired gene also depends on the accuracy of the DNA polymerase to assemble the oligonucleotides into longer DNA sequences. However, with the appearance of high-fidelity DNA polymerases, additional errors introduced by DNA polymerases have dramatically decreased. Thus, errors accumulated during oligonucleotide production are the principal obstacle to efficient gene synthesis. Error rates of 1-10 errors per kilobase of DNA have been reported in different studies (Binkowski, 2005; Hoover & Lubkowski, 2002; Hughes *et al*., 2011; Xiong *et al*., 2004). It was also observed that error frequency increases as the length of an oligonucleotide increases (Xiong *et al*., 2004).

2.4.1.3.3.1. Removing errors from oligonucleotides

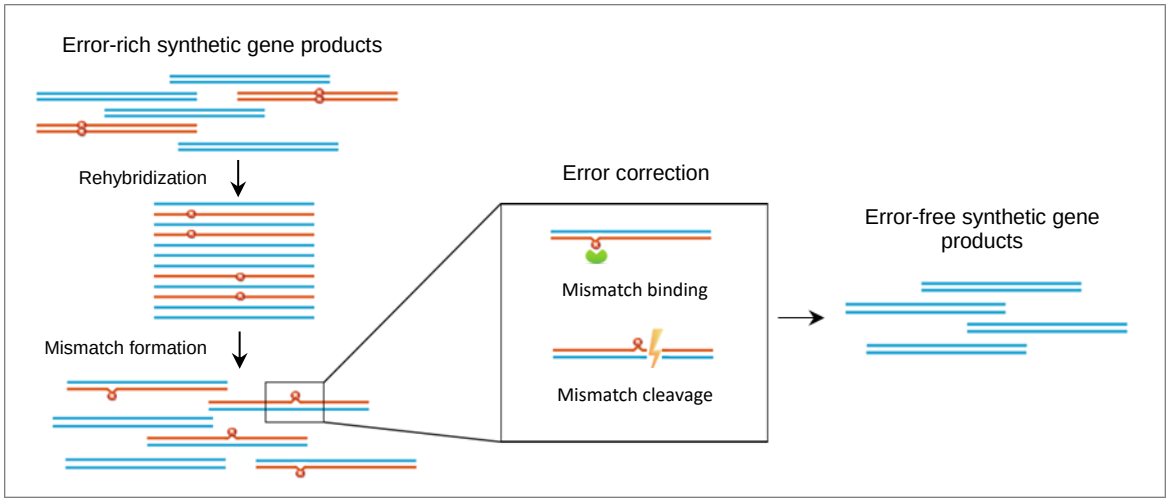
As described above, the quality and purity of the oligonucleotides used in gene synthesis is the principal factor affecting the accuracy of the procedure. Several gene synthesis methods use oligonucleotides that were previously purified to remove deletions and/or insertions. Oligonucleotide purification is commonly realized by size exclusion purification using high performance liquid chromatography (HPLC) or polyacrylamide gel electrophoresis (PAGE). These methods allow the elimination of >90% of the impurities (mostly insertions, deletions and truncations) before gene assembly and, as a result, the number of errors identified in synthetic gene can be reduced. However, the use of additional purification steps is costly, labour intensive and low-throughput. Previous studies show that length of oligonucleotides used for gene assembly influences error rates of the final products (Xiong *et al*., 2004; Young & Dong, 2004), since the oligonucleotide length can vary from 40 bp to sizes over 100 bp. The use of short oligonucleotides is usually the best choice although in some cases may improve costs due to an increase in the number of oligonucleotides and overlap regions required (Xiong *et al*., 2008).

2.4.1.3.3.2. Error removal from synthetic genes

Sequence errors that remain in oligonucleotides will be carried over during the assembly process and will consequently accumulate in the synthetic full-length genes. Therefore, the selection of a synthetic gene without errors often requires expensive and time consuming cloning and sequencing steps. Considering that errors identified in synthetic genes can be the cause of reading frame shift and/or loss of functions from target protein, it was necessary to develop strategies to reduce the number of errors in synthetic DNA constructs and consequently increase the efficiency of the selection process of an error-free synthetic gene. Most of the current approaches to remove errors from synthetic DNA constructs are based on

the use of DNA mismatch recognition proteins which have the ability to recognize mismatches between two DNA strands. However, if errors identified in synthetic genes are presents in both DNA strands, this means that no mismatches between DNA strands will be formed and, consequently, no DNA correction. After gene synthesis, resulting genes are denatured to favour a re-annealing reaction forming DNA hetero-duplexhetero-duplexes. The random re-association of the polynucleotide chains allows the formation of DNA mismatches through hybridization of incorrect bases with the corresponding correct bases in the reverse-complementary strand. Subsequently, hetero-duplexhetero-duplexes that contain mismatches can then be recognized and/or removed by using mismatch binding proteins or mismatch-cleavage enzymes (Figure 2.10).

Figure 2.10| Illustration of the mismatch-based error correction approach used in gene synthesis.



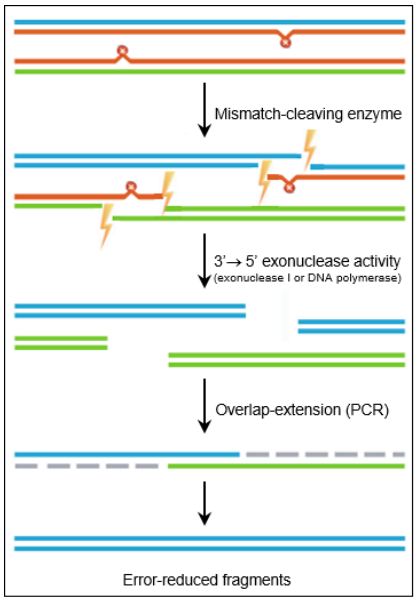
The formation of DNA hetero-duplexhetero-duplexes is performed by heat denaturation and then reannealing to allow the rehybridization of correct (blue lines) with mutant (red lines) strands and subsequent formation of random mismatches. The mismatches are then removed by two different error correction methods using either mismatch-binding proteins or mismatch-cleavage enzymes to enrich error-free synthetic genes sequences. (Adapted from Ma *et al*., 2012).

Mismatch binding proteins have the capacity to selectively bind to mismatches generated by incorrect re-hybridization of DNA strands. The MutS protein is a typical mismatch binding protein as it recognizes and binds to a variety of mispaired bases and small single-strand loops. For this reason, MutS can be used as an error-removal agent in gene synthesis methods. MutS is part of the MutHLS mismatch repair machinery of *E. coli* and it is used together with MutH and MutL proteins to locate, bind, and cleave mismatches duplexes *in vivo*. A thermostable version of MutS from *Thermus aquaticus* was used by Carr *et al*. (2004) to remove gene products containing mismatches from synthetic genes. In this study, the thermostable MutS protein was immobilized on a solid support, where it recognizes and binds to hetero-duplexhetero-duplexes DNA sequences that contain mismatches, and a gel-mobility shift

assay leads to the separation of mutated DNA sequences from unbound homoduplexes (mostly error-free). This method was shown to reduce errors by more than 15 fold relative to conventional gene synthesis techniques and has an error rate of only one per 10 kb (Carr, 2004). An alternative MutS-based error-correction method to enrich correct genes was presented by Binkowski *et al*. (2005) and it is called “consensus shuffling”. Hetero-duplexHetero-duplexes containing mismatches are cleaved with an endonuclease and the resulting short fragments are then subjected to MutS column filtration. Short fragments containing mismatches are captured by immobilized thermostable MutS, whereas error-free fragments are eluted and reassembled into correct full-length sequence by PCR assembly. The authors demonstrated that two iterations of consensus shuffling reduce the error rate by 3.5-4.3 fold, being the error rate of only one error per 3500 bp (Binkowski, 2005). This method provides several advantages over direct MutS filtration of full-length sequences, as it is more effective for longer DNA sequences and it tolerates more errors in the starting polynucleotide chains (Ma *et al*., 2012).

Mismatch cleavage proteins correspond to a group of specific endonucleases that identify and cleave mismatch sites generated by hybridization between correct and incorrect DNA strands. This group of endonucleases includes: resolvases, such as phage T4 endonuclease VII, T7 endonuclease I and *E. coli* endonuclease V, and single-strand specific nucleases, such as S1 nuclease from *Aspergillus orizae*, P1 nuclease from *Penicillum citrinum*, mung bean nuclease and CEL nuclease from celery (Ma *et al*., 2012). The ability of these endonucleases to cleave hetero-duplexhetero-duplexes DNA at the mismatch sites made them a versatile tool in error removal for gene synthesis methods, and different assays based on their activity have been developed in order to enrich error-free sequences in previously assembled DNA constructs. A general error correction reaction using mismatch cleavage enzymes is based on the previous denaturing and re-annealing reaction to produce mismatch chains and the recognition and cleavage of DNA sequences with mismatches (Figura 2.11). Cleaved fragments are removed either by size exclusion or repaired by an exonuclease, such as *E. coli* exonuclease I, or a proofreading DNA polymerase, which has 3’-5’ exonuclease activity. After cleavage, DNA fragments are extended using repaired overlapping fragments into full-length sequences (Fuhrmann, Oertel, Berthold, & Hegemann, 2005; Ma *et al*., 2012). The efficacy of several mismatch cleavage proteins has been explored to recognize and cleave single-base mismatches using different assays designed for elimination of failure DNA sequences from a pool of assembled DNA sequences.

Figure 2.11| Schematic representation of the principle of error removal when are used mismatch-cleaving enzymes.



Mismatch-specific enzymes recognize DNA mismatches and cleave reannealed hetero-duplexhetero-duplexes in both strands, near to DNA mismatch, generating short overhangs. The short overhangs are immediately dissociated at the temperature reaction. The single-stranded extensions that contain the mismatch bases are degraded by a single-strand-specific 3'-5'-exonuclease, e.g. by *E. coli* exonuclease I or the corresponding activity of proofreading DNA polymerase. Full-length genes with a reduced number of errors are recovered by overlap assembly PCR (Fuhrmann *et al*., 2005; Ma *et al*., 2012). Figure adapted from Ma *et al*., 2012.

Fuhrmann *et al*. (2005) explored the activity of three resolvases - phage T4 endonuclease VII, T7 endonuclease I and *E. coli* endonuclease V - to cleavage double-strand DNA containing single mismatched base pairs in the bacterial chloramphenicol-acetyltransferase (*cat*) gene. The data revealed that these endonucleases identify and cleave with high efficacy synthetic DNA sequences containing mutations. Moreover, T4 and *E. coli* endonucleases reduced the occurrence of mutations in synthetic genes about 4-fold when compared with the non-treated controls. On the other hand, T7 endonuclease I showed to be less efficient in detection of errors than T4 and *E. coli* endonucleases (Fuhrmann *et al*., 2005). Contradictory results have been reported by Tsuji *et al*. (2008) which indicated that T7 Endonuclease I has much better efficiency than T4 endonuclease VII and *E. coli* endonuclease V (Tsuji & Niida, 2008). Single-strand specific (sss) nucleases have also been used in assays to distinguish between correct DNA sequence and undesirable DNA sequences that contain mismatches. S1, P1 nucleases from fungi, the mung bean nuclease and the CEL nuclease from plant are the most widely used nucleases. Several studies have demonstrated that S1 and P1 nucleases are strongly specific to AT-rich regions (Yeung, Hattangadi, Blakesley, & Nicolas, 2005), although S1 nuclease seems incapable of recognizing single base mismatches (Silber & Loeb, 1981). CEL endonuclease, an orthologue of S1 nuclease isolated from celery, has the ability to cleave with high specificity all single base pair mismatches (B. Yang *et al*., 2000). It is not inhibited by high GC content, and can cut DNA hetero-duplexhetero-duplexes at neutral pH if the mismatches

are base substitutions, insertions or deletions anywhere from 1 to 12 nucleotides. Additionally, CEL nuclease nicks a DNA strand at the 3'-end of the base mismatch and it is able to cleave DNA molecules with multiple mismatches (B. Yang *et al*., 2000). Moreover, CEL nuclease is reported to display lower mismatch bias and higher digestion efficiency when compared with T4 endonuclease VII, T7 endonuclease I and *E. coli* endonuclease V (Tsuji & Niida, 2008). The broad substrate specificity and low undesired activity of CEL nuclease makes it the most promising candidate for error correction in synthetic genes (Ma *et al*., 2012; Saaem, Ma, Quan, & Tian, 2012; B. Yang *et al*., 2000). In Chapter 4, we report the development of an error correction assay using a mismatch-cleaving enzyme, phage T7 endonuclease I, which was integrated in gene synthesis method. This endonuclease was used to reduce the number of mutations in synthetic genes with more than 500 bp.

2.4.1.4. Fusion tags to improve recombinant protein expression in *E. coli*

Escherichia coli is the most widely used host system for the production of heterologous proteins due to specific features, such as (1) rapid growth at high cell density on inexpensive substrates; (2) short times between cells generations; (3) cells do not require specialized equipment for cultivation; (4) easy manipulation due to its well characterized genetics; (5) and the availability of a large number of molecular tools and protocols, such as cloning vectors with different N- and C-terminal tags, engineered strains and cultivation approaches. In addition, this prokaryotic host system often provides high yields of recombinant proteins (Mancia & Love, 2011; Rosano & Ceccarelli, 2014). However, there are some difficulties associated with the expression of specific heterologous proteins in this host. The principal difficulty is the production of inclusion bodies, usually as a consequence of high levels of recombinant protein expression and inappropriate conditions for correct protein folding. Further problems include toxicity, low levels of expression, protein degradation and production of non-functional protein derivatives (Peleg & Unger, 2012). Thus, several different strategies have been developed to produce suitable amounts of recombinant proteins in soluble and biologically active form in *E. coli* by increasing both yield and solubility. Quite often eukaryotic recombinant proteins need post-translational modifications in order to become active and/or adopt their proper structure, which does not operate if protein expression is directed to the cytoplasm of *E. coli*. In order to solve this problem, several *E. coli* strains have been genetically modified to allow the introduction of several post-translational modifications, such as disulfide-bond formation in the cytoplasm by providing oxidizing conditions due to mutations in thioredoxin reductase (*trxB*) or/and glutathione reductase (*gor*) genes in AD494 and Origami™ (Novagen) strains, or by co-production of DsbC proteins in SHuffle® strains (Novagen) (Berkmen *et al*., 2012; Bessette, Aslund, Beckwith, & Georgiou, 1999; Derman, Prinz, Belin, & Beckwith, 1993) . In addition, some recombinant proteins may be toxic for *E. coli* cells as they may carry out a negative function in the host cell leading to abnormal proliferation and homeostasis of the

microorganism (Rosano & Ceccarelli, 2014). Thus, in these cases, basal levels of expression need to be tighter controlled and different approaches have been developed with this goal such as the addition of glucose to the growth medium (Studier, 2005), co-expression of T7 lysozyme (Moffatt & Studier, 1987), which is performed in BL21pLysS and BL21pLysE strains, using low copy number plasmids, such as the pETcoco vectors (Novagen) (Wild, Hradecna, & Szybalski, 2002) or directing expression to the periplasm or media, where expression yields are usually lower.

Several studies have demonstrated that fusion of recombinant peptides and polypeptides with highly soluble protein partners to form a chimeric protein, promotes both protein yield and solubility (Terpe, 2003). The most popular fusion partners used to enhance protein solubility are the glutathione S-transferase (GST), the maltose binding protein (MBP), the N-utilization substance protein A (NusA), the thioredoxin (Trx), the small ubiquitin-like modifier (SUMO) and the ubiquitin (Ub) (Table 2.2). More recently, other fusion partners have been proposed, such as the 8-kDa calcium binding protein Fh8 from the parasite *Fasciola hepatica* (Costa, Almeida, Castro, Domingues, & Besir, 2013).

Table 2.2| Principal properties of the most common protein fusion tags used in recombinant protein expression in *Escherichia coli*.

Tag	Residues/ Size (KDa)	Matrix/ Elution	Comments	Ref.
Fh8	69/ 8.0	An affinity tag must be added (usually His-tag)	Small tag; Ca^{2+}-dependent binding to phenyl-Sepharose	1
Trx	109/ 11.7	An affinity tag can be added (usually His-tag)	-	2
SUMO (Smt3)	101/ 11.6	An affinity tag must be added (usually His-tag)	Cleavage by SUMO Protease 1	3
Ub	128/ 14.73	An affinity tag must be added (usually His-tag)	-	4
DsbA	208/ 23.1	An affinity tag must be added (usually His-tag)	Introduces disulfide bonds; enables protein solubilization in the periplasm or in the cytoplasm	5
DsbC	216/ 23.4	An affinity tag must be added (usually His-tag)	Isomerization of disulfide bonds; enables protein solubilization in the periplasm or in the cytoplasm	6
GST	211/ 26.0	Glutathione-agarose/ glutathione	GST dimerization and glutathione elution may affect fusion protein properties	7
MBP	396/ 42.5	Cross-linked amylase/ maltose	Large tag; Matrix compatible with nonionic detergents and high salt, but not reducing agents	8
NusA	495/ 54.87	An affinity tag must be added (usually His-tag)	Large tag, may affect properties of fusion protein	9

References: 1) Costa *et al*. (2013); 2) LaVallie *et al*. (1993); 3) Butt *et al*. (2005); 4) (Baker, 1996); 5) Collins-Racie *et al*. (1995); 6) Nozach *et al*. (2013); 7) Smith & Johnson (1988); 8) di Guana *et al*. (1988); 9) Davis *et al*. (1999).

The mechanism used by fusion tags to promote expression and solubility of adjacent proteins remains unclear. It was proposed that fusion of a stable or conserved molecule to an insoluble recombinant protein may stabilize and promote proper folding of the recombinant protein. In addition, fusion tags may act as a nucleus of folding (Englander, 2000). For example, it was shown that MBP possesses an intrinsic chaperone-like activity (Kapust & Waugh, 1999; Raran-Kurussi & Waugh, 2012). MBP has emerged as the preferred solubility tag for a range of diverse proteins and it was used for the efficient production of recombinant disulfide-rich venom peptides (Klint *et al*., 2013). The data revealed that directing MBP to the periplasm played a crucial role in the improvement of disulfide-bonds formation, which are required for the biological activity of venom peptides. Fusion technology was also shown to be a useful tool for the protection of the recombinant proteins from degradation; fusion can promote translocation of the "unwanted" recombinant protein to different cellular compartments avoiding the exposition to proteases. For instance, MBP may be involved in the translocation of proteins to the membrane (Nikaido, 1994). Other proteins such as disulfide isomerases (e.g. DsbA and DsbC) have also been proposed as fusion partners. These fusion partners have proved to enhance solubility and proper folding of proteins in the non-reducing periplasmic environment or in the cytoplasm, if expressed without an efficient signal peptide (Baneyx, 1999; Nozach *et al*., 2013). Several studies have reported that the DsbC fusion tag is a good choice for improving the expression levels of disulfide-rich peptides (Nozach *et al*., 2013; Saez, Nozach, Blemont, & Vincentelli, 2014). Recently, two transmembrane small proteins, Yoag and YkgR, which are orientated with their N-termini in the cytoplasm and their C-termini in the periplasm, were used for fusion expression of two disulfide bond-rich peptides. High levels of correct folded peptides were obtained using these two new fusion tags (Chang *et al*., 2015). Other fusion partners, such as NusA, SUMO, Trx and Ub require an affinity tag, such as the poly-histidine for protein purification. Otherwise, MBP and GST can serve to purify the recombinant protein by affinity chromatography, as MBP binds to amylose-agarose and GST to glutathione-agarose, respectively (Rosano & Ceccarelli, 2014). There are several studies comparing the effects of various fusion tags on protein yields and levels of soluble recombinant protein obtained (Braun *et al*., 2002; Dyson, Shadbolt, Vincent, Perera, & McCafferty, 2004; Hammarström, Hellgren, van den Berg, Berglund, & Härd, 2002; Shih *et al*., 2002). However, the inconsistency of the data from these comparative studies suggests that each protein or class of proteins has unique optimal conditions and that fusion tags vary greatly in efficiency. The correct choice of an appropriated fusion tag should take into account its size, since this parameter plays a critical role in the total yield of the target protein, as well as its effects on the tertiary structure or biological activity of the fused protein (Balbás, 2001). In addition, removal of the fusion tag must be considered when structural or biochemical studies on the target recombinant protein are required (Balbás, 2001). Cleavage of fusion tags can be performed through two different ways: chemically (Chong *et al*., 1997) or by using an enzymatic approach

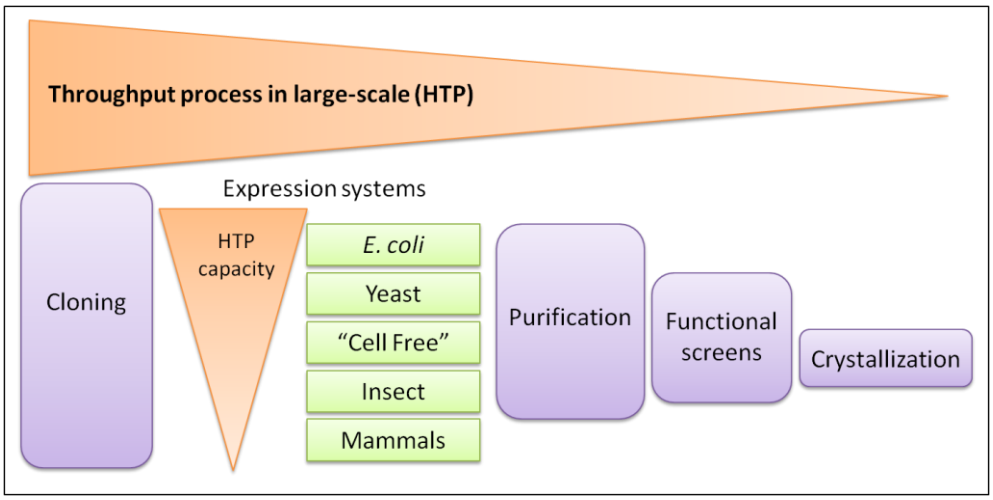
(Jenny, Mann, & Lundblad, 2003). When the first method is selected, the fusion tag is removed by treatment of the fusion protein with a chemical reagent in harsh conditions (Hwang, Pan, & Sykes, 2014). The second strategy involves the insertion of a unique amino acid sequence that is susceptible to cleavage by a highly specific protease, such as Tobacco Etch Virus (TEV) protease, factor Xa, thrombin protease and the SUMO protease (Blommel & Fox, 2007; Jenny *et al*., 2003; Satakarni & Curtis, 2011). In opposition, Koehn & Hunt (2009) demonstrated that after tag removal, some proteins become unfolded and, consequently, precipitated (Koehn & Hunt, 2009). In addition, complete cleavage rarely occurs leading to a reduction of the target protein yield (Baneyx, 1999). In summary, a fusion partner should ideally not compromise the tertiary structure and biological activity of the fused protein, be easy to remove without affecting protein structure after removal, allow the implementation of simple purification procedures and be applicable to a range of proteins (Terpe, 2003). To fulfil these requirements, new tag-protein fusion systems are constantly emerging, allowing improving the efficacy of soluble protein production, which is particularly relevant for high-throughput protocols. In Chapter 5, we report a comprehensive study that compares the efficiency of different fusion tags for the production of recombinant venom peptides. This work allowed to select the best fusion tag to ensure high levels of stable and active recombinant peptides in *E. coli*.

2.4.2. High-throughput (HTP) methodologies for protein research

Advances in sequencing technologies have generated enormous amounts of genetic information derived from both genome and metagenome projects. The challenge of the post-genomic era is to develop new ways to functionally analyse large amounts of data derived from sequencing projects. Computational analysis and genome annotation using several search alignment tools (BLAST) (Altschul *et al*., 1997) attempt to assign functions based on protein homology for the majority of predicted proteins encoded by sequenced genomes. Presently, detection of amino acid sequence similarities to proteins of known function allows the annotation of 40-70% of novel genome sequences by homology (Eisenberg *et al*., 2000). Nevertheless, the fraction of proteins without functional annotation remains large. Genomic and proteomic analysis offer the promise of assigning a biological function to all the proteins encoded by the genome of an organism. However, the optimal utilization of genomic sequence data requires rapid and efficient methods for the generation of expression clones and the evaluation of protein production, thus leading to the rapid protein characterization and structure determination. Thus, high-throughput methodologies have emerged as an alternative to change the paradigms from the past, leaving behind the *modus operandi* of analysing one gene-one protein and moving on to a time where the simultaneous analyses of multiples genes/proteins is a reality. HTP approaches of the post-genomics era require the implementation of novel methods for gene synthesis, cloning, protein expression, purification and detection that allow working with large numbers of genes and proteins and, at the same

time, to analyse the enormous amounts of data that are generated. Even for laboratories which are studying a single protein target, these steps are usually expensive and time-consuming. Solutions to overcome these problems have emerged from structural genomics projects, which use a standard experimental workflow and a “funnel” approach (Figure 2.12).

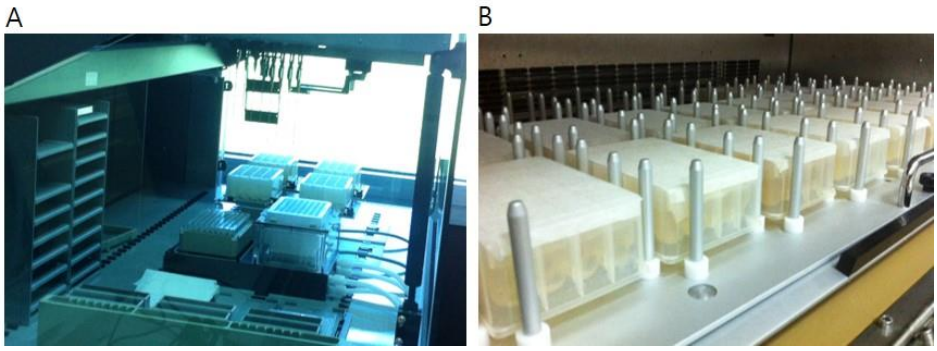
Figure 2.12| The typical “funnel” scheme in high-throughput structural studies applied to protein research.



The throughput of the entire process decreases from cloning to crystallization. The last step (not showed) involves progression from crystallization to structure determination which usually cannot be fully automated. Throughput also diminishes as the complexity of the expression systems increase. Adapted from Mancia and Love (2011).

Each stage of the process needs to be optimized on parameters like throughput, automation, speed, cost-effectiveness and scalability (Mancia & Love, 2011). Implementation of an HTP platform for gene synthesis, cloning and expression of hundreds to thousands of targets, requires significant modifications in traditional gene synthesis, cloning and expression protocols. Novel methodologies for the automated generation of synthetic genes and protein expression clones using microwell plates, multichannel pipettors and, in some cases, commercially available liquid handlers (Figure 2.13) have been reported (Bruni & Kloss, 2013). This usually involves protocol miniaturization employing 24-, 96- or 384-well plate formats (Abdullah, Joachimiak, & Collart, 2009; Bruni & Kloss, 2013; Dieckman, Gu, Stols, Donnelly, & Collart, 2002; Scheich, Sievert, & Büssow, 2003).

Figure 2.13| Automated solutions compatible with HTP platform for gene synthesis of target genes.



Panel A: Liquid handling workstation and panel B: HTP incubator shaker with 24-deep well plates that are frequently used for plasmid propagation or protein expression in large scale.

2.4.2.1. HTP methods for gene synthesis

The emergent field of synthetic biology is generating insatiable demands for synthetic genes, which far exceed existing gene synthesis capabilities. To achieve this goal, it is crucial to develop a gene synthesis technology that allows to synthesise multiple *de novo* DNA constructs of any size or sequence using rapid, accurate, high-throughput, and cost-effective protocols. Methods for *de novo* chemical synthesis of DNA have been refined over the last years providing improved protocols for the production of recombinant, mutated, or completely novel DNA sequences. In general, the entire gene synthesis process involves five different steps: (1) sequence optimization and oligonucleotides design, (2) oligonucleotides synthesis, (3) gene assembly using PCR-based strategies, (4) sequence verification and error removal, and (5) synthetic gene product preparation for downstream applications. Only one of these steps, namely the oligonucleotide synthesis, has been automated and new instruments were developed in order to create a HTP gene synthesis approach that increases the quality and throughput of synthetic genes production (Tian *et al*., 2009). Several methods for DNA oligonucleotides synthesis based on automated synthesisers have been developed with a throughput ranging from 1 to 1536 nucleic acids produced simultaneously (Cheng, Chen, Kao, Kao, & Peck, 2002; Horvath, Firca, Hunkapiller, Hunkapiller, & Hood, 1987; Lashkari, Hunicke-Smith, Norgren, Davis, & Brennan, 1995; Sindelar & Jaklevic, 1995). Furthermore, DNA microarrays and microfluids devices for DNA synthesis have recently been the focus of attention to create an inexpensive and high-throughput next generation method for oligonucleotide production (Tian *et al*., 2009). DNA microarray synthesis technologies utilize the surface of a silicon chip or glass slide, which contains numerous spots, to elongate the oligonucleotide chain through a variety of different mechanisms, including photolithography, inkjet printing, electrochemical array and microfluidics. These mechanisms allow to control whether or not a phosphoramidite monomer will be coupled to the growing oligonucleotide chain on a particular spot during each synthesis cycle (Gao, Gulari, & Zhou, 2004). Other

oligonucleotide synthesis technologies were also developed to decrease cost and to increase the throughput of the gene synthesis process (Tian *et al*., 2009).

One important drawback of current protocols for the production of synthetic genes is the absence of a defined and robust approach integrating all steps of gene synthesis. In general, present methods in synthetic biology use separate and not integrated protocols for gene and oligonucleotide design, nucleic acid assembly, error correction and cloning. Currin *et al*. (2014) provided the first example of an integrated gene synthesis method that requires fewer steps than those described previously. This integrated method includes different steps and contemplates gene design, oligonucleotides design and synthesis, PCR-based gene assembly and error removal based on endonuclease cleavage. The method, called SpeedyGenes, was recently presented as a rapid methodology for the synthesis of protein libraries and enables the efficient synthesis of large and multiple genes (Currin, Swainston, Day, & Kell, 2014).

In summary, current methods developed to produce synthetic genes display a low-throughput and high-cost. However, numerous improvements have recently been realized leading to further decreases in cost and increases in throughput, automation and platform integration. This will further promote the establishment of the *de novo* gene synthesis as a routine and standard method for molecular biology and genetic engineering. In addition, HTP gene synthesis will be a powerful tool for creating new biological products that can potentially transform biomedical research. This thesis reports the development of an innovative HTP platform for gene synthesis that was used to produce thousands of genes encoding venom peptides (Chapter 3). The creation of this integrated method aimed to revolutionize the gene synthesis methods through the development of an efficient, robust and cost-effective HTP platform that can support the use of recombinant venom peptides for drug discovery.

2.4.2.2. HTP methods for gene cloning

Several methods for the fast and cost-effective cloning of large numbers of open reading frames (ORFs) into expression vectors have been developed in the last years. These usually involve Ligation Independent Cloning (LIC) and do not require the use of restriction enzymes or DNA ligases. Marsischky and LaBaer (2004) defined an optimal HTP cloning method as a method ease to use, reliable, flexible and inexpensive. Based on this principle, several properties should be considered by these innovative methods: (1) the transfer of DNA cloned from master clones to expression plasmids must be 100% (or almost) efficient, conservative, thereby avoiding mutations, and should result in the correct orientation of ORFs, (2) the validation procedure for the cloned products should be simple (ideally only a single clone for each target gene should be sequenced), (3) the cloning system should be able to support the transfer of genes into virtually any type of expression vector, (4) the addition of fusion tags or cloning related sequences to an ORF should be minimal since they often affect the subsequent expression and crystallization of the recombinant protein (Marsischky & LaBaer, 2004). In

addition, an HTP cloning method should be independent of the sequence of the target gene and a single PCR amplification should be sufficient for cloning into different vectors. As described above, conventional cloning methods based on DNA cleavage by restriction endonucleases and the subsequent ligation using DNA ligases ("cut and paste") are, in this respect, unsatisfactory, because they are relatively inefficient, time-consuming and labour-intensive. LIC methods attempt to overcome the limitations of ligation-based cloning and are based on a recombination reaction that occurs between the insert and the destination vector. These technologies include various systems such as the Gateway (Hartley, Temple, & Brasch, 2000), the Creator (Colwill *et al*., 2006), the MAGIC (Li & Elledge, 2005), the In-Fusion (Berrow, Alderton, & Owens, 2009) or the sequence- and ligation-independent cloning (SLIC) that relies on homologous recombination (Li & Elledge, 2007). Other LIC technologies use complementary single-strand overhangs on the vector and insert, which allow cloning by base complementation without the need of a ligation step. These methods include, for instance, the LIC method based on T4 DNA polymerase (Tachibana *et al*., 2009). In general, sequence-dependent methods are less convenient in HTP procedures because they require unique and specific sites in both the insert and the vector. Thus, more flexible sequence-independent cloning methods are preferred. The preparation of insert and/or vector DNA fragments was described as a critical point in LIC methods, since these additional steps are often time-consuming and usually employ expensive enzymes (Quan & Tian, 2009).

HTP cloning requires efficient methods for the selection of the correct clones, which is usually a labour-intensive and often low efficient task due to parental vector background. Thus, many positive selection vectors have been developed to improve the efficacy of this step. Successful cloning using these vectors usually involves a change of the host phenotype, which can be achieved either by the inactivation or the replacement of a gene marker (e.g. the lethal gene ccdB) through insert cloning (Haag & Ostermeier, 2009; Hu, Zhang, Li, & Wang, 2010).

Despite the reasonable effectiveness of the currently available LIC systems for HTP cloning, it is still required to develop alternative protocols for the rapid and efficient creation of expression clones. Thus, several research groups have contributed to improve the efficiency of current methods applied to HTP cloning leading to more robust, reliable and cheaper protocols (Bryksin & Matsumura, 2010; Geertsma & Dutzler, 2011; J. Yang, Zhang, Zhang, & Luo, 2010).

2.4.2.3. HTP methods for protein expression and purification

There are different expression systems available for the high-throughput production of recombinant proteins. *Escherichia coli* is the simplest host for recombinant protein expression and has been the most readily adapted expression system to the HTP format. Recombinant protein expression in *E. coli* has proved to be fast, cost-effective and scalable (Chambers, 2002; Mancia & Love, 2011). In bacteria, the most common promoter used to control gene

expression is the T7 promoter and it is usually regulated by an operator where the repressor (lacIQ) binds, allowing a tight regulation of gene expression. The inducer is usually lactose or quite often the non-hydrolysable lactose analogue isopropyl β-D-1-thiogalactopyranoside (IPTG) and when bound to the repressor, leads to a structural change in the protein lacIQ that involves its displacement from the DNA chain and allows the formation of the hybrid T7/lac promoter. The T7/lac system provides extra levels of expression control, to address problems related with the production of toxic proteins (i.e. ribosome destruction, cell death and plasmid or expression instability) (Koehn & Hunt, 2009). Nevertheless, *E. coli* is not suitable to the production of large and complexes eukaryotic proteins that require post-translational modifications and eukaryotic expression systems arise as a good choice for the production of biologically active complex polypeptides. However, in this case, the throughput of protein expression usually decreases, also accompanied by low yields and high costs (Hartley, 2006). Several mammalians (Heyman *et al*., 1999), yeast (Schuster, 2000) and insect cells (Albala *et al*., 2000; Coleman *et al*., 1997) have been used as expression systems compatible with HTP protocols. Cell-free expression systems have also recently emerged as an alternative expression system for recombinant protein expression, especially to produce toxic and insoluble proteins. This system has shown to be useful in HTP approach allowing expressing proteins directly from PCR-generated transcripts with no need to generate expression vectors or manipulate cells in culture (Chambers, 2002).

Development of HTP procedures usually involve the cultivation of *E. coli* cells in auto-induction media, first described by Studier (2005). These media allow the simultaneous induction of expression of multiple recombinant proteins under the control of a T7/*lac* promoter and contain specific components that after an initial period of tightly regulated growth allow fully automated induction of target protein at high optical densities. Thus, usage of these media allows regulated induction of protein expression with no need to monitor cells growth and induction with IPTG or analogues (Studier, 2005b). Even using 96-well plates, high cell densities can be achieved (Lesley, 2001). Different components of complex media are reported to support or suppress growth to high cell density of a wide range of *E. coli* strains with different nutritional requirements (Studier, 2005). Lactose can support growth of *E. coli*, but several restrictions are reported when it is used as a carbon and energy source for high-level production of target protein (e.g. lactose's catabolism leads to the production of galactose which is not used as a carbon source by BL21, the most common *E. coli* protein expression strain) (Studier, 2005). Thus, cultures growing in media containing glucose and lactose will utilize all glucose before starting to metabolize lactose (following a catabolite repression mechanism). To prevent lactose metabolism, a control of gene expression relies on the binding of the CAP-cAMP complex to the promoter required for the transcription of the lac operon. The presence of glucose is associated with the presence of the complex. Once glucose concentration increases inside the cell, the cAMP decreases, as the amount of the complex also decreases. In this

situation, the complex does not bind to the lac promoter and the lac operon is turned off (Epstein, Rothman-Denes, & Hesse, 1975). When lactose is present, it binds to the lac repressor protein (encoded by lacI), making it unable to bind to the operator. Therefore, lactose acts as the inducer of the transcriptional expression controlled by the T7 promoter (Horton, Lewis, & Lu, 1997). There are some auto-induction media solutions sold commercially; being the most used the Overnight Express™ Autoinduction system (Novagen). Auto-induction media provide a convenient solution for HTP procedures since the *E. coli* cells only have to be inoculated and grown to saturation, without monitoring their growth and without adding any additional expression inductor. In addition, this system allows reaching higher yields of soluble protein production, typically several-fold higher than those obtained by standard IPTG induction (Studier, 2005b). SDS-PAGE and Coomassie-blue staining can be used for a primary and rapid analysis of levels of protein expression. However, the complete information of the expressed protein will be acquired after cell lysis, protein purification, and analysis of recombinant protein biological activity.

Immobilized Metal Affinity Chromatography (IMAC) is one of the most robust and efficient methods for protein purification. Current recombinant expression systems involve engineered specific affinity tags in the recombinant proteins that allow the rapid and efficient implementation of purification protocols. Purification based on histidine tags (His-tag) is a universal solution for purifying proteins in parallel and in a single-step. Thus, His-tags (which bind to immobilized divalent metal ions, Ni^{2+}) are widely used for protein purification and usually provide high levels of purity that are satisfactory for most downstream applications. Several metal chelating resins with high affinity for His-tags have been adapted to the 96-well format protocol for protein purification using automated systems (Chambers, 2002). In general, IMAC automatable protocols involve multiple steps. After cultivation, usually performed using deep-well plates, cells are collected by centrifugation and are lysed using either mechanical disruption or chemical lysis. Mechanical disruption may be achieved using a deep-well microplate-horn sonicator commercially available. Alternatively, chemical lysis has been successfully applied in many HTP protocols since it is fast, robust and less labour-intensive than the mechanical lysis (Chambers, 2002). There are many reagents for bacterial cell walls disruption and most of them include lysozyme and treatments with DNA nucleases. In addition, non-ionic and zwitterionic detergents can be used for nondenaturing lysis (Lesley, 2001). BugBuster® (available from Novagen), B-PER® (from Thermo Scientific, Pierce Protein Biology Products) and NZY Bacterial Cell Lysis Buffer (NZYTech, genes & enzymes) are three examples of detergent-based reagents available commercially. Automatable protein purification is initiated when the crude cell lysate is mixed with the nickel charged resin to capture the recombinant protein, and then transferred to 96-well filtration plates. The purification system requires the use of vacuum pressure to allow the passage of cell extracts

and wash buffers through the plate columns (Scheich, Sievert, & Büssow, 2003). Depending on the expression level and the volumes used, roughly 1 µg to 6 mg of target protein can be purified from small-scale *E. coli* cultures (1-10 mL). This protein yield is frequently enough for expression analysis and for implementation of initial functional studies (Chambers, 2002). Automation in structural and functional assays has reduced the amount of concentrated protein, being 1-10 mg usually sufficient for these procedures (Lesley, 2001).

2.5. Venoms as therapeutics

Animal venoms represent a huge untapped resource of bioactive molecules, which may be of diagnostic or therapeutic benefit for public health. In recent times, venom peptides have been the subject of intense scientific investigation, in part due to the recent interesting "biologic" molecules for drug development (Sébasti Dutertre *et al*., 2015). Reticulated peptides from venoms were defined as a novel class of biologics (Escoubas & King, 2009). Venom peptides often show higher potency, greater target-specificity, higher resistance to degradation, and lower immunogenicity than small-molecular drugs. Thus, peptides have emerged as an important class of therapeutics in modern drug discovery pipelines (King, 2011; Lewis & Garcia, 2003). For example, venom peptides have already been applied in the treatment of several human diseases, including hypertension, diabetes, chronic pain, ischemic stroke, AIDS and cancer (Lewis & Garcia, 2003; Vetter *et al*., 2011). Furthermore, peptides are valuable research tools to explore the physiological functions of many human receptors and to discover the biological mechanisms underlying disease.

2.5.1. Venom-based drug discovery

The traditional approach towards venom-based drug discovery is based on activity-guided fractionation, where venoms are screened in assays against targets of therapeutic interest, then "hit venoms" are chromatographically fractionated and individual fractions re-screened in order to isolate peptides with bioactivity. These peptides are then sequenced via a combination of Edman degradation, tandem mass spectrometry (MS/MS) techniques, and venom-gland transcriptomics (Sébasti Dutertre *et al*., 2015). These methods have proved to be useful for acquiring information related with the biochemical properties of toxins and to discover new drug leads. However, these approaches are usually laborious and typically require large amounts of venom. Furthermore, classical methods applied to venom research also display low throughput and are only adapted for big animals that produce large volumes of venoms; collection of sufficient quantities of venoms from small/rare animals is often not practical nor sustainable. In order to expedite the process of discovery and to decrease amount of venom required for bioactivity-guided fractionation, an integrated approach that combines transcriptomic and proteomic methodologies together with powerful bioinformatics tools has been eagerly required (Sébasti Dutertre *et al*., 2015; Prashanth, Lewis, & Dutertre, 2012). It

is expected that these novel strategies will deliver a maximum of information from limited sample volumes. Recent improvements in the fields of transcriptomics, proteomics and bioinformatics have positive impacts in drug discovery in particular when these three platforms are integrated towards a more efficient discovery strategy. Thus, a multidisciplinary approach termed Venomics emerged to complement the conventional bioassay-guided strategies.

2.5.1.1. Transcriptomics

Conventional genomic and transcriptomic methods have been extensively applied to study the genes that encode for the venom bioactive peptides and proteins. These methods are based on the construction of cDNA libraries followed by Sanger sequencing of expressed sequence tags (ESTs). However, the amount of information obtained through Sanger sequencing is not complete and this approach suffers from low coverage of the transcriptome; hence does not present the entire view of venom components. The recent advances in next-generation sequencing allowed the development of highly innovative HTP platforms for DNA sequencing, such as Illumina (Illumina) and the 454 pyrosequencing (Roche). These have allowed the study of an entire transcriptome or genome in a rapid and cost effective manner. Thus, transcriptomics analysis allows achieving a more comprehensive view of venom peptides. This information should be validated by proteomics data using mass spectrometry, as described below. In summary, several studies have demonstrated the suitability of using transcriptomics approaches to generate valuable information enabling the rapid discovery of novel venom components.

2.5.1.2. Proteomics

The traditional approach to obtain the amino acid sequence of peptide drug leads is based on the use of automated Edman degradation. With an increase on the number of peptides to analyse, the sequencing by Edman degradation proved to be more expensive, low throughput and requires large sample volumes. Recently, advances in the field of proteomics and improvements in mass spectrometry instrumentation have made proteomic studies an important approach to unravel the true complexity of venoms. Animal venoms are predominantly composed of peptides and are thus highly suitable to an investigative approach based on mass analysis. Due to the high complexity of most venoms it is necessary to combine several techniques of liquid chromatography with mass spectrometry, such as high-performance liquid chromatography (HPLC) and liquid chromatography (LC). Also, Matrix-Assisted Laser Desorption-Ionisation (MALDI) Time of Flight (TOF) mass spectrometry and Electrospray Ionisation (ESI) mass spectrometry have gained popularity in venom-based drug discovery laboratories. Mass spectrometry techniques have been widely used in venom profiling and whole venom fingerprinting. Presently, the proteomic analyses of venoms have evolved to the *de novo* sequencing of peptides using mass spectrometry approaches.

De novo sequencing allows to confirm the identification of venom components where databases do not exist, and also obtaining full sequences of venom components for which small amounts do not permit the application of Edman sequencing. In addition, this method is faster and cost-effective, while requiring small amounts of venom material when compared with traditional methods (Escoubas *et al*., 2008; Prashanth *et al*., 2012). Two different approaches have been developed for venom *de novo* sequencing. The first one, termed "bottom-up", is the most common and is suitable for sequencing of high mass peptides of >3,5 kDa. Following the bottom-up approach, peptides are purified followed by reduction and alkylation of the disulfide bonds, after which they are digested by an endopeptidase such as trypsin. The second approach is termed "top-down". The entire peptide is fragmented in the gas-phase without enzymatic digestion. Within the context of a venom-based drug discovery pipeline, proteomics studies are an integral part of venom analysis, representing major tools for the identification of novel therapeutic molecules.

2.5.1.3. Bioinformatics

Bioinformatics is an important field at every step of venom-based discovery. It provides the tools to improve the efficiency of venom research programs by organising large datasets in searchable databases, while providing efficient methods to facilitate data analysis and compare different datasets. Recent advances in proteomics and transcriptomics provide large amounts of venom peptide data, increasing the need for efficient data management and to develop novel purpose-built bioinformatics tools. Broadly, the function of bioinformatics in venom research can be separated in two distinct fields: data management and data analysis. Data management refers to the compilation of data from different sources in the form of a database, which provides researches with relevant information in a rational form. Databases combining venom data usually provide an overview of toxin properties, emphasizing the venomous species or the type of toxins that have higher potential to become drugs, and suggest which venomic fields are still unexplored. Examples of toxin databases are: ConoServer (http://www.conoserver.com) and ArachnoServer (http://www.arachnoserver.com) that contain expert annotation on the sequences and 3D structures of cone snails and spiders, respectively. Other example of a venom database is the animal toxin database (ATDB, http://protchem.hunnu.edu.cn/toxin/) that displays information from several databases. ATDB combines detailed ontologies which describe the function of toxins and target ion channels (He *et al*., 2010). In addition, data analysis refers to the computational tools that allow researchers to analyse large amount of raw data. Thus, bioinformatics tools are necessary for the successful establishment of venom research programs. In most cases, these tools are designed *de novo* and contain specifications that are specific for each step of venom-based drug discovery process. Many bioinformatics tools are

integrated within databases and, for instance, are used to analyse sequences, derive phylogenetic relations or 3D structure characteristics (Quentin & Craik, 2015).

The modern era of venom-based drug discovery has been established as an integrated approach that combines the activity-guided fractionation strategy with proteomics and transcriptomics approaches, and bioinformatics tools. Thus, it is possible to screen crude venom and identify “hit venoms” with bioactivity using the traditional strategy, and to discover the sequence of venom components by new “omics” approaches, which require small amounts of venom material. Additionally, a rapid and efficient production system is required for production of sufficient quantities of synthetic peptides, which are used for complete structural, functional and *in vivo* characterization, and to screen for novel drug leads.

2.5.2. Pharmaceutical use of venom peptides

The medical value of toxins has been known from ancient times. For example, the use of scorpion and snake venoms are used in folk remedies, and in Western and Chinese traditional medicine (Koh, Armugam, & Jeyaseelan, 2006). Later on, toxins were isolated, biologically characterized and used as drugs. The true reason for this is that toxins are highly refined by the evolution process, up to the point where every molecule is endowed with valuable pharmacological properties. The venom-based drug discovery began in the 1970s with the development of the blockbuster drug captopril (CAPOTEN®) an antihypertensive synthetic molecule that structurally and functionally mimics peptides discovered in the venom of the Brazilian viper *Bothrops jaracaca* (Cushman & Ondetti, 1991). Subsequently, more venom-derived drugs have been approved by the United States Food and Drug Administration (FDA). Prialt® (zinoconotide) is one of the most successful venom drugs that corresponds to the synthetic version of ω-conotoxin MVIIA, a peptide isolated from the venom of the marine snail *Conus magus*. Prialt® was approved by the FDA in 2004 as an analgesic for the treatment of chronic pain (Miljanich, 2004). The most recent venom drug is Exenatide (Byetta®), a synthetic version of exendin-4 from the saliva of the Gila monster lizard. It is a peptide agonist of the glucagon-like peptide receptor that has been approved as an adjuvant in the treatment of adults with Type 2 diabetes (Bray, 2006). Toxins are used in different medications (Table 2.3) and in a large variety of diagnostic assays related with the haemostatic system (Takacs & York, 2014). Toxins as drugs are either used as a natural toxin purified directly from crude venom (e.g., batrotoxin), synthetic version of the natural toxin (e.g., exenatide, ziconotide), or as a peptide (e.g., eptifibatide) or nonpeptide (e.g., tirobifan, captopril) peptidomimetic of the natural toxin (King, 2011; Takacs & York, 2014). Currently, the majority of venom-derived drugs are derived from the venoms of various species of vipers (Viperidae), the European medicinal leech (*H. medicinalis*), Gila monster (*Heloderma suspectum*) and marine snails (*Conus magus*). These data show that venom-derived drugs have had a huge impact on medicine. For example, two of the three available agents in the platelet glycoprotein inhibitor class of drugs

are snake venom derived (Takacs & York, 2014). A large number of toxins and toxin-derived compounds are still in various stages of development from the experimental phase to clinical trials (King, 2011), such as the novel neurotoxin Brachyin that was identified in the venom of the spider *Brachypelma albopilosum*. This neurotoxin is a reticulated peptide of 41 amino acids that showed significant analgesic effects in mice models (Zhong *et al*., 2014). TM-601, currently in Phase II human trials, is a modified form of the scorpion peptide chlorotoxin that selectively targets receptors on glioma cells – a diffuse form of brain cancer, without binding to healthy surrounding neurons (Mamelak & Jacoby, 2007). Other biomedical applications have been attributed to toxins, such as cosmetic, anti-venoms and biopesticides.

Table 2.3| Drugs derived from animal venom toxins.

Drug name	Species origin	Mechanism of action	Indication
Captopril (CAPOTEN®)	Jararaca	Angiotensin-converting enzyme inhibitor	Hypertension, cardiac failure
Enalapril[a] (VASOTEC®)	Jararaca	Angiotensin-converting enzyme inhibitor	Hypertension, cardiac failure
Exenatide (BYETTA®)	Gila monster	Glucagon-like peptide-1 receptor agonist	Type 2 diabetes mellitus
Exenatide (BYDUREON®)	Gila monster	Glucagon-like peptide-1 receptor agonist (extended release)	Type 2 diabetes mellitus
Ziconotide (PRIALT®)	Magician's cone snail	$Ca_v2.2$ channel antagonist	Management of severe chronic pain
Bivalirudin (ANGIOMAX®)	European medicinal leech	Reversible direct thrombin inhibitor	Anticoagulant
Lepirudin (REFLUDAN®)	European medicinal leech	Binds irreversibly to thrombin	Anticoagulant
Desirudin (IPRIVASK®)	European medicinal leech	Selective and near-irreversible inhibitor of thrombin	Prevention of venous thrombotic events
Tirofiban (AGGRASTAT®)	Saw-scaled viper	Antagonist of fibrinogen binding to GPIIb/IIIa receptor	Acute coronary syndrome
Eptifibatide (INTEGRILIN®)	Pigmy rattlesnake	Prevents binding of fibrinogen, von Willebrand factor, and other adhesive ligands to GPIIb/IIIa receptor	Acute coronary syndrome
Batroxobin (DEFIBRASE®)	Common lancehead or Brazilian lancehead	Cleaves Aα-chain of fibrinogen	Acute cerebral infarction; unspecific angina pectoris; sudden deafness
Platelet gel (PLATELTEX-ACT®)	Common lancehead	Cleaves Aα-chain of fibrinogen	Gelification of blood for topical applications in surgery
Fibrin sealant (VIVOSTAT®)	Brazilian lancehead	Cleaves Aα-chain of fibrinogen	Autologous fibrin sealant in surgery
Thrombin-like enzyme	Chinese moccasin or Siberian pit viperor	Fibrinogenase	"Antithrombotics"; "defibrinating" agent for the treatment and prevention of thromboembolic diseases"
Hemocoagulase (REPTILASE®)	Common lancehead or Jararaca or Brazilian lancehead	Cleaves Aα-chain of fibrinogen; factor X and/or prothrombin activation	Prophylaxis and treatment of haemorrhage in surgery
Medicine leech therapy	Medicinal leech	Inhibit platelet aggregation and the coagulation cascade	Skin grafts and reattachment surgery

[a] Enalapril is a later-generation derivatives of captopril. Adapted from Takacs & York, 2014.

2.6. VENOMICS project

VENOMICS project was an innovative European Union funded project dedicated to the exploration of venomous animal biodiversity for public health. The core objective of VENOMICS was to recreate *in vitro* collections of venom peptides that can be used as a resource for high-throughput screening, helping to more efficiently isolate novel drug leads. Thus, this approach aimed to create "synthetic venoms" in the laboratory to be used in drug discovery programs. For this purpose, a novel research paradigm was established combining “omics” technologies in a high-throughput workflow. This project was conducted in a large scale, with the goal of reproducing the diversity of venom peptides via high-throughput recombinant expression, synthesis and refolding of reticulated peptides. VENOMICS aimed to develop an automated platform for the high-throughput production of reticulated peptides, combining refolding tools, to allow the construction of peptide banks equivalent to “natural venoms” in a scale compatible with the amount of sequence information generated by the proteomics and transcriptomics workflows. The project pipeline was initiated with a source of 200 different venomous species from which biological samples were extracted. Then, the venom gland transcriptome and venom proteome were analysed in order to create a 25,000 peptide sequences database. From this sequences bank, a unique library of 6,000 venom peptides was aimed to be produced by recombinant expression (5,000 peptides) and chemical synthesis (1000 peptides), depending on the size and the presence of PTMs. As final goal, the peptide library was used in the identification of novel therapeutic leads using specific assays which wished to mirror important metabolic pathways.

The project involved a European consortium composed of seven specialized partners from universities, small and medium companies and institutes that had expertise in different areas, such as animal venom research, high-throughput protein production, proteomic analysis, transcriptomic analysis, molecular biology, drug development and project management. Each partner had indispensables skills, knowledge and physical resources that combined it into a unique multidisciplinary consortium. The VENOMICS project was funded by the seven research framework programme (FP7) of the Research and Technological Development branch of the European Commission for four years, from 2011 to 2015. The results of VENOMICS project are not yet fully known. However, this project developed a successful research program that combines consolidated tools to explore and understand animal venoms in a scale never reached before. In addition, the database of peptide sequences and the constructed synthetic peptides library represent a huge improvement, with an extraordinary value, for the research venom and scientific community. Thus, VENOMICS will have a major impact on public health issues by offering both innovative receptor-targeted drugs as well as novel therapeutic avenues for a number of current unmet medical needs. This thesis was developed under the VENOMICS project, namely in the development of an innovative high-throughput platform for the simple, fast, efficient and cost-effective production of hundreds of

synthetic genes encoding venom peptides. Chapter 6 reports the production of 4992 synthetic genes encoding venom peptides that were produced using the novel HTP gene synthesis platform developed in Chapter 3.

2.7. Objectives

The work presented here aims to develop a high-throughput gene synthesis platform to produce hundreds to thousands of synthetic genes encoding disulfide-rich venom peptides. This platform will constitute an innovative tool to explore the enormous potential of venom peptides, in particular to accelerate the discovery of novel venom-based drugs. The novel protocols could be used to optimize and produce any DNA sequence that encodes a protein of interest and are in agreement with the development of innovative molecular biology products that the Biotechnology sector pursues. Specifically, the main goals of this project may be summarized as follows:

- To develop a high-throughput platform for the production of synthetic genes encoding venom peptides, using rapid, accurate and cost-effective protocols (Chapter 3).
- To create an error correction system based on mismatch recognition proteins to reduce the number of mutations identified in synthetic genes, as presence of errors in artificial genes are the major bottleneck of current gene synthesis methodologies (Chapter 4).
- To develop a codon optimization algorithm adapted to design nucleic acids encoding venom peptides for expression in *E. coli* (Chapter 5).
- To develop novel strategies to enhance the levels of recombinant protein production in *E. coli*, including the development of an efficient vector to express disulfide-rich peptides and novel mechanisms to remove fusions partners using TEV protease (Chapter 5).
- To produce 5000 synthetic genes encoding venom peptides using the HTP gene synthesis platform developed in previous chapters (Chapter 6).

3. DEVELOPMENT OF A GENE SYNTHESIS PLATFORM FOR THE EFFICIENT LARGE SCALE PRODUCTION OF SMALL GENES ENCODING ANIMAL TOXINS

Ana Filipa Sequeira[1,2], Joana L.A. Brás[2], Catarina I.P.D. Guerreiro[2], Renaud Vincentelli[3] and Carlos M.G.A. Fontes[1,2]

[1] Centro Interdisciplinar de Investigação em Sanidade Animal (CIISA) - Faculdade de Medicina Veterinária, Universidade de Lisboa, Avenida da Universidade Técnica, 1300-477 Lisboa, Portugal; [2] NZYTech Genes & Enzymes, Campus do Lumiar, Estrada do Paço do Lumiar, Edifício E, r/c, 1649-038 Lisboa, Portugal; [3] Unité Mixte de Recherche (UMR) 7257, Centre National de la Recherche Scientifique (CNRS) – Aix-Marseille Université, Architecture et Fonction des Macromolécules Biologiques (AFMB), Campus de Luminy, 163 Avenue de Luminy, 13288 Marseille CEDEX 09, France.

Adapted from a submitted manuscript.

Abstract

Gene synthesis is becoming an important tool in many fields of recombinant DNA technology, including recombinant protein production. *De novo* gene synthesis is quickly replacing the classical cloning and mutagenesis procedures and allows generating nucleic acids for which no template is available. In addition, when coupled with efficient gene design algorithms that optimize codon usage, it leads to high levels of recombinant protein expression. Here, we describe the development of an optimized gene synthesis platform that was applied to the large scale production of small genes encoding venom peptides. This improved gene synthesis method uses a PCR-based protocol to assemble synthetic DNA from pools of overlapping oligonucleotides and was developed to synthesise multiples genes simultaneously. This technology incorporates an accurate, automated and cost effective ligation independent cloning step to directly integrate the synthetic genes into an effective *Escherichia coli* expression vector. The robustness of this technology to generate large libraries of dozens to thousands of synthetic nucleic acids was demonstrated through the parallel and simultaneous synthesis of 96 venom genes. Large scale recombinant expression of synthetic genes encoding eukaryotic toxins will allow exploring the extraordinary potency and pharmacological diversity of animal venoms, an increasingly valuable but unexplored source of lead molecules for drug discovery.

3.1. Introduction

Synthetic biology, an interdisciplinary branch of biology, is quickly becoming one of the most attractive areas of research and development thanks to the recent developments in gene synthesis technology. In combination with intelligent gene design methods, gene synthesis is

emerging as a valuable tool to support recombinant protein expression. *De novo* gene design allows optimizing codon usage to the recombinant host system thus promoting the effective operation of the cellular translational machinery. In addition, when the nucleic acid template is not available, gene synthesis allows creating *de novo* DNA molecules. This is of extreme importance considering the exponential growth of genomic and metagenomic information and the current limitations in using this highly useful sequence information due to the lack of tangible DNA.

In recent years, a variety of gene synthesis methodologies have been developed based on the assembling of oligonucleotides into complete genes. Early approaches advanced to synthesise nucleic acids used the enzymatic ligation of pre-formed duplexes of phosphorylated overlapping oligonucleotides (Ashman, Matthews, & Frank, 1989). Subsequently, self-priming PCR (Hayashi *et al*., 1994), PCR assembly (Hoover & Lubkowski, 2002), Polymerase chain assembly (PCA) (Stemmer, Crameri, Ha, Brennan, & Heyneker, 1995) and template-directed ligation (Strizhov *et al*., 1996) were described as efficient concepts for *de novo* gene synthesis. Recently, methods based on a two-step approach were reported for the production of long DNA sequences. Examples of these methods are the PCR-based thermodynamically balanced inside-out technology (TBIO) (Xinxin Gao *et al*., 2003), the two-step total gene synthesis method (Young & Dong, 2004) that combines both dual asymmetrical PCR (DA-PCR) and overlap-extension (OE-PCR), the PCR-based two-step DNA synthesis (PTDS) (Xiong *et al*., 2004) and PCR-based accurate synthesis (PAS) (Xiong *et al*., 2006). Lately, improvements in PCR-based gene synthesis methods, as exemplified by the development of the improved PCR synthesis (IPS) and the simplified gene synthesis (SGS) protocols (Gordeeva, Borschevskaya, & Sineoky, 2010; G. Wu *et al*., 2006), have been described and incorporate significant simplifications over earlier strategies. SGS uses oligonucleotides of 40 nucleotides (nt) in length and 18-20 nt of overlap region, which are assembled in a unique PCR-assembly reaction leading to the direct construction of the full-length DNA molecule. The simplicity of this protocol combined with its relative low cost, as there is no requirement for phosphorylation or purification of the oligonucleotides, are a solid base for the development of even more effective PCR-based methods. However, major drawbacks persist and effective improvements need to be implemented in current synthetic protocols to allow their translation to a large scale. One of the major bottlenecks of current gene synthesis protocols consists on the quality of the oligonucleotides used for nucleic acid assembly. It is known that all current gene synthesis methods accumulate errors in the final synthetic molecules. Sequence errors usually derive from the incorporation of imperfect synthetic oligonucleotides or result from low fidelity associated with enzymatic assembly. Current oligonucleotide synthesis methods produce sequences that are often prematurely terminated, or comprise internal mutations (error rates range from 1 to 10 mutation per kilobase (kb)) (Binkowski, Richmond, Kaysen, Sussman, & Belshaw, 2005). In addition, chemical synthesis of DNA molecules usually not only involve

moderate to high error rates but also high costs. Moreover, the chemical synthesis of a desired gene also depends on the accuracy of the DNA polymerase used to assemble the oligonucleotides in a final DNA sequence. Therefore, DNA errors are inevitable and it is often necessary to remove the incorrect synthetic DNA molecules using enzymatic methods (Binkowski, 2005; Carr, 2004). Improvements in oligonucleotide quality, error correction and DNA polymerase efficacy are thus urgently required.

Conventionally, PCR-based gene synthesis is employed to produce a single gene at a time. Thus, development of automated platforms that effectively generate large libraries of nucleic acids is urgently needed. The different steps leading to a single PCR-assembly strategy need to remain simple, accurate and robust when extended to the assembly of multiple genes simultaneously. To develop large-scale methods, many factors that affect the efficiency of gene assembly, such as DNA polymerases performance or oligonucleotide concentration and quality require optimization. This work describes different approaches carried out to optimize current gene synthesis protocols. The data was integrated to develop a novel platform, which was applied to efficiently synthesise and clone a large number of nucleic acids encoding venom peptides. This automated platform can be translated to the rapid generation of complex gene libraries encoding different families of relevant biotechnologically valuable proteins and peptides.

3.2. Materials and Methods

3.2.1. Gene design

The original purpose of this research was to optimize protocols for the synthesis of small genes encoding eukaryotic peptides for expression in *Escherichia coli*. Three genes of different lengths (A: 290 bp, B: 260 bp, and C: 329 bp) were selected to develop these studies. Gene A encodes the *alpha-elapitoxin-Nk2a* toxic protein isolated from the Naja kaouthia venom, and genes B and C encode two different venom peptides of unknown function. Venom genes were designed by back-translating the peptide sequence and optimizing codon usage for high levels of expression in *E. coli*. Gene design was performed using ATGenium codon optimization algorithm developed in the scope of this thesis. DNA sequences of the three genes are presented in supplementary (see Table S3. 1 in Annex). ATGenium uses a Monte Carlo repeated sampling algorithm to randomly select a codon for each amino acid according to their frequency defined in a codon frequency lookup table for the selected host system. The lookup table applied to design gene sequences comprises the codon usage for *E. coli*, which includes codons used preferentially in highly expressed or average native *E. coli* genes. Gene design maximized stable mRNA molecules, minimized the presence of repeated sequences (no more than 5 identical nucleotides) and avoided the appearance of *E. coli* regulatory sequences such as promoters, activators or operators. In addition, codon adaptation index (CAI) value was set to be higher than 0.8 and guanine-cytosine (GC) content was set to vary between 40% and

60%. The sequence generated by back-translating each peptide is checked to ensure that it adheres to the specified factors described above. If the gene sequence contains avoided sequences, the gene design process is repeated up to 1000 times until an acceptable DNA sequence is generated. Moreover, ATGenium was integrated in a bioinformatics software that was prepared to generate multiple gene sequences simultaneously to improve the throughput of the process. Thus, using a Microsoft Excel interface, this software needs as an input a variable number of protein sequences for back-translation. Thus, this bioinformatics tool allowed the automation of the gene synthesis pipeline facilitating gene design when multiple protein sequences are used as template.

3.2.2. Oligonucleotides and purification

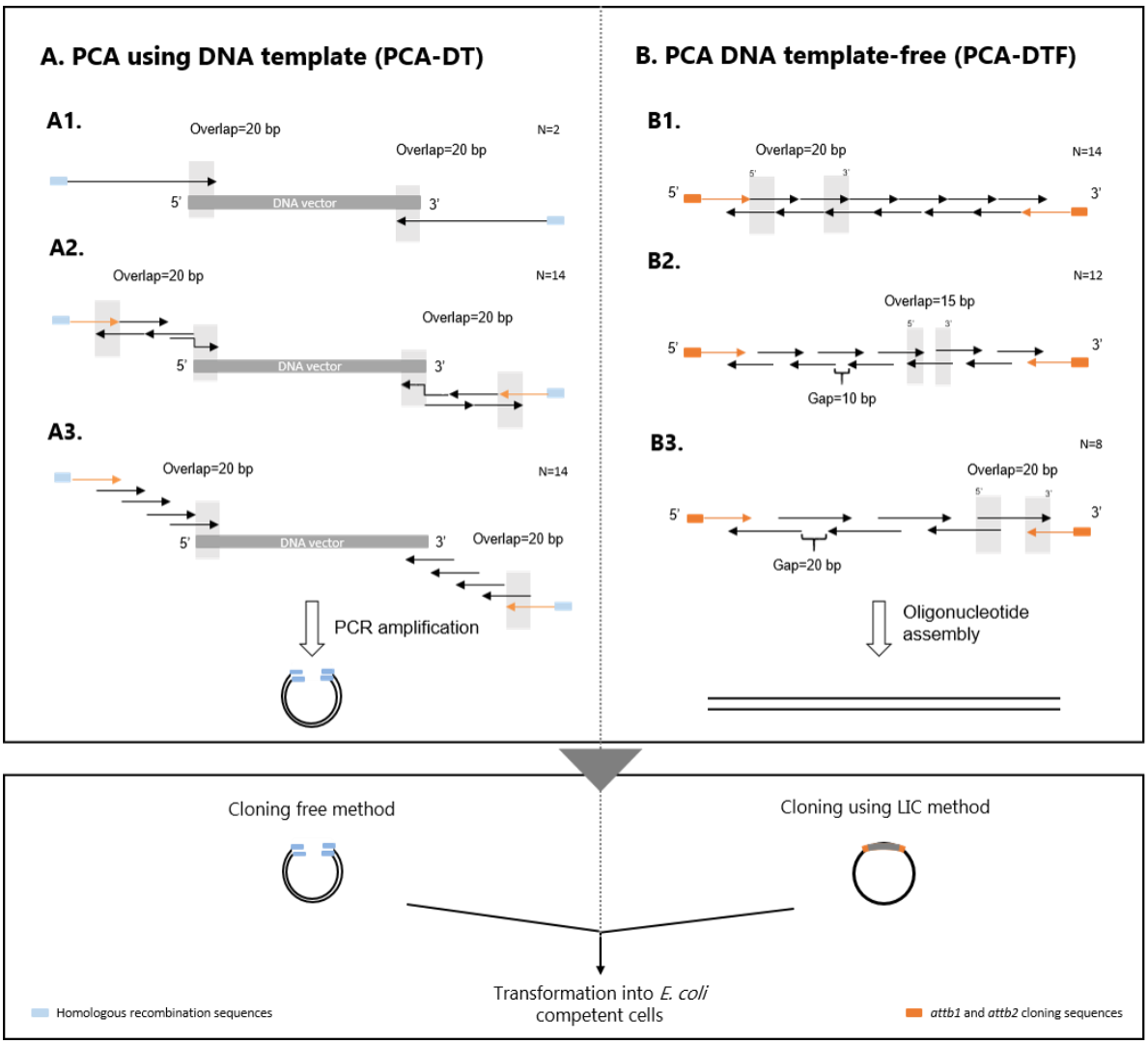
Oligonucleotides were synthesised by three different suppliers (A, B and C) using the smaller scale available with standard desalting. Reverse-phase cartridge and reverse-phase HPLC purifications were also tested for oligonucleotides originated from supplier C. Each oligonucleotide was reconstituted in 10 mM Tris-HCL (pH 8.5) to a 100 µM final concentration and kept at -20 ºC. Oligonucleotides were used individually or in mixtures at concentrations described below.

3.2.3. Primer design

We have developed a dedicated software (NZYOligo designer) for the large scale design of primers for gene synthesis. As the ATGenium software this program can work simultaneously with multiple DNA sequences. The algorithm used for oligonucleotide design allows defining primer lengths, gap regions, overlapping regions by defining start and end positions, and introducing engineered 5´ or 3´-end sequences. NZYOligo designer uses a Microsoft Excel interface and requires as an input multiple gene sequences that need to be artificially produced. The DNA sequence of each gene is used as the template to design the assembly oligonucleotides by dividing the entire sequence into overlapping primers with defined lengths. NZYOligo designer allows automating the gene synthesis pipeline, particularly the primer design steps and it ensures an accurate design of multiples oligonucleotides simultaneously. The external oligonucleotides, termed outer primers, correspond to the external forward and reverse primers. For strategy A (see below), the outer primers include a complementary sequence of 15 bp to promote plasmid re-circulation through homologous recombination (Figure 3.1, A). Following strategy B (see below) the outer primers contained an additional sequence at the 5'-terminus of both forward and reverse primers for cloning into the cloning vector through a Ligation-Independent Cloning protocol (LIC) (Figure 3.1, B). Internal oligonucleotides (termed inner primers) are usually present in higher number than outer primers. Depending on the gene assembly strategy used, the oligonucleotides were designed

with gaps between adjacent primers and contain 15-20 bp overlap regions. The sequence of all oligonucleotides used in this study is displayed in supplementary (see Table S3.2 in Annex).

Figure 3.1| Schematic diagram representing the two different approaches used for gene synthesis: A. Polymerase chain assembly using DNA template (PCA-DT), and B. Polymerase chain assembly DNA template-free (PCA-DTF).



Each oligonucleotide is represented as an arrow; black arrows correspond to internal (inner) oligonucleotides while external (outer) oligonucleotides are denoted as orange arrows. Each strand of the desired gene is dissected into two or more oligonucleotides that are amplified (A) or overlap extended (B) by a DNA polymerase. A1-Two long oligonucleotides were designed with a 20 bp overlap region with the cloning vector; A2-Two sets of 40 bp oligonucleotides containing the 5'-end and 3'-end sequences of the desired gene were assembled by PCR; A3 – Successive oligonucleotides with 40 bp in length were designed with a 20 bp overlaps regions between adjacent oligonucleotides to construct the full-length of cloning vector containing the desired gene. The PCR products from A1, A2 and A3 strategies were inserted in *E. coli* cells through homologous recombination. Strategy B corresponds to a single-step PCR assembly step that does not require template DNA. B1 – Gapless 40 bp oligonucleotides containing 20 bp overlap regions between primers were used to assemble the synthetic gene; B2 – 40 bp oligonucleotides with 15 bp overlaps between forward and reverse primers, with gaps of 10 bp were mixed to produce the gene of interest; B3 – The synthetic gene was synthesised by mixing a 60 bp overlapping oligonucleotides containing gaps of 20 bp. After overlap extension by DNA polymerase, gene fragments from B1,

B2 and B3 were cloned into a vector using a ligation-independent cloning method. All outer primers contain a vector complementary region at the 5'-end (highlighted in orange rectangles) or a 16 bp complementary sequence (highlighted in blue rectangles) to facilitate the cloning reaction.

3.2.4. Novel strategies to synthesise small genes

In order to develop an efficient, cost-effective and low error rate strategy to produce synthetic genes with reduced size, two different strategies to construct optimized DNA sequences were initially explored: (A) Polymerase chain assembly using DNA template (PCA-DT) and (B) Polymerase chain assembly DNA template-free (PCA-DTF) (Figure 3.1, A and B, respectively). PCA-DT was developed to decrease the time involved in traditional gene synthesis methods since this method combines both synthesis and cloning in a single reaction. In step 1, two long (A1) or a pool of small oligonucleotides containing the gene sequence (A2 and A3) are mixed with the cloning vector and then assembly proceeds using a DNA polymerase in a typical cyclic temperature reaction (Figure 3.1, A) (Table 3.1). In step 2, the product of the PCR amplification combining both the newly synthesised gene and the vector are used to transform *E. coli* cells. PCA-DTF strategy is based on previously reported methods used to produce synthetic genes in a single PCR reaction (Stemmer *et al*., 1995; G. Wu *et al*., 2006; Xiong *et al*., 2006). All oligonucleotides (inner and outer) are pooled together and assembled in a single polymerase chain reaction. The outer primers are used in a higher concentration than inner primers to ensure the construction of full-length sequence of synthetic gene. Different approaches for oligonucleotide design were tested (Figure 3.1, B1, B2 and B3). The PCA-DTF method requires a subsequent ligation-free cloning step to insert the synthetic gene into the cloning vector.

Table 3.1| Gene assembly strategies used to produce synthetic gene A.

Gene assembly method	DNA template	Number of primers	Primer size (nt)	Cloning method
PCA-DT_A1	yes	2	135	Homologous recombination
PCA-DT_A2	yes	14	27-40	Homologous recombination
PCA-DT_A3	yes	12	35-40	Homologous recombination
PCA-DTF_B1	no	14	35-40	Gateway system
PCA-DTF_B2	no	12	30-40	Gateway system
PCA-DTF_B3	no	8	35-60	Gateway system

(A) PCA-DT

Three different strategies based on PCA-DT method were used to synthesise gene A (290 nt). Two, fourteen and twelve oligonucleotides were designed to synthesise full-length gene A following strategies A1, A2 and A3, respectively. For A1 strategy, the gene sequence was dissected in two long oligonucleotides of 135 nt, including a 20 nt overlap with the cloning vector. To produce the synthetic gene using A2 strategy, fourteen oligonucleotides of 20-40 nt

with a 20 bp overlap region between forward and reverse primers and no gaps between adjacent oligonucleotides were designed. The length of twelve oligonucleotides used in A3 was 35-40 nt and the overlapping region between successive oligonucleotides was 20 bases. All outer primers contain an additional 15-bp homologous sequence to facilitate the homologous recombination reaction before *E. coli* transformation. Plasmid pNZY28 (NZYTech, Ldt) was used as a cloning vector and was linearized with EcoR V restriction enzyme during 2 hours at 37ºC in a heating block. A typical digestion was performed in 100 µL containing 2 µg of plasmid DNA and 50 units of EcoR V restriction enzyme. Linear plasmid DNA was purified using silica-based columns, eluted in 50 µL elution buffer and diluted to a final concentration of 20 ng/µL. The synthesis of gene A using the strategy A1 was initiated with the addition of the two outer primers at a final concentration of 200 nM to 20 ng of digested pNZY28 vector. For strategies A2 and A3, outer and inner primers were used at a final 800 nM and 30 nM concentration, respectively. The PCR reaction was carried out with 200 µM dNTPs and 2.5 units of Pfu Turbo DNA polymerase (Agilent Technologies). The PCR conditions were 30 cycles at 95ºC for 50 s, 50ºC for 50 s and 72ºC for 3 min. The final cycle was followed by an additional 10 min at 72ºC to ensure complete extension of the 3,099 bp gene product (pNZY28 vector: 2,886 bp + gene A: 213 bp). PCR amplification products were analysed by agarose gel electrophoresis.

(B) PCA-FDT

To synthesise gene A based on strategies B1 and B2 of PCA-DTF, fourteen and twelve oligonucleotides with 40 nt in length and 15 or 20 nt end-overlaps between consecutive oligonucleotides, respectively, were designed (Figure 3.1, B1 and B2). Strategy B3 used larger oligonucleotides with 60 nt including 20 nt overlaps (Figure 3.1, B3). Outer primers include an additional ligation independent sequence (27 nt on the forward primer and 32 nt on the reverse primer), in order to allow ligation-independent cloning and a 18 nt encoding the Tobacco Etch Virus (TEV) protease. Oligonucleotide assembly was performed as described above using *Pfu* Turbo DNA polymerase (Agilent Technologies). Assembly reaction was subjected to one cycle of initial denaturation at 95ºC for 5 min, followed by 26 cycles of denaturation at 95ºC for 30 s; annealing at 55ºC for 30 s; and extension at 72ºC for 30 s. PCR amplification products were column purified, as described above, and cloned into pDONR201 vector using Gateway® technology (ThermoFisher Scientific) (see below).

3.2.5. Optimization of PCR conditions for successful gene synthesis protocol

Efficacy and accuracy of DNA polymerases and the quality and concentration of primers are two critical parameters known to influence gene synthesis. In addition, annealing temperatures and times used for denaturation, annealing and extension during PCR may affect nucleic acid yields. To optimize these parameters that are critical to PCR assembly, two genes (B and C) were synthesised using gene synthesis strategy B3. Six and eight oligonucleotides with 58-60

bp and a 20 bp gap between primers were used to synthesise genes B and C, respectively. The sequences of overlapping oligonucleotides used to produce genes B and C are presented in Table S3.3 in Annex. The effect of each PCR parameter was single tested and the remaining components of PCR reaction were fixed in the standard conditions described above. Four DNA polymerases were selected for these studies: KOD Hot Start DNA polymerase (EMD-Millipore), Q5® Hot Star High-Fidelity DNA polymerase (New England Biolabs), Pfu Turbo DNA polymerase (Agilent Technologies) and Taq DNA polymerase (Sigma-Aldrich). PCR was developed in a 26-cycle reaction. Denaturation was performed at 95ºC for 30 s for Pfu Turbo DNA polymerase and Taq DNA polymerase, 95ºC for 16 s for KOD Hot Start DNA polymerase and 98ºC for 10 s for Q5® Hot Start High-Fidelity DNA polymerase. Annealing occurred at 60ºC for 10 s for KOD Hot Start DNA polymerase and 60ºC for 30 s for Q5® Hot Star High-Fidelity DNA polymerase, Pfu Turbo DNA polymerase and Taq DNA polymerase. Finally, extension was performed at 70ºC for 3 s for KOD Hot Start DNA polymerase, 72 ºC for 30 s for Pfu Turbo DNA polymerase and Taq DNA polymerase, and 72ºC for 15 s for Q5® Hot Star High-Fidelity DNA polymerase. Different overlapping oligonucleotide concentrations were tested. Gene assembly was performed using inner oligonucleotides at a final concentration of 10, 20 and 30 nM. Outer primers were used at final concentrations of 200, 600, 800 and 1000 nM. In addition, the final concentration of dNTPs was set to vary between 0.1 and 0.5 mM. Finally, different PCR profiles were tested. Thus, PCRs were performed at five different annealing temperatures (from 50ºC to 62ºC). Furthermore, 24 PCR programs, which included different times of denaturation, annealing and extension in 22, 24 and 26 cycles, were tested. Final configuration of each PCR program used here is presented in supplementary (see Table S3.4 in Annex).

3.2.6. Cloning and sequencing

After PCR assembly, resulting nucleic acids were purified, inserted into a suitable vector and the integrity of each gene was confirmed by DNA Sanger sequencing. Synthesised genes produced by strategy A were ligated into pNZY28 cloning plasmid using the homologous recombination machinery present in *E. coli* cells. Gene products from strategies B1, B2 and B3 which contained *attb1* and *attb2* sequences at its 5' and 3'- ends, respectively, were cloned into pDONR201 vector (ThermoFisher Scientific) using the Gateway® cloning system (ThermoFisher Scientific). Cloning reaction mixtures were used to transform *E. coli* DH5α competent cells. For each transformation, one bacterial colony was inoculated and grown in liquid LB medium supplemented with 100 µg/mL of ampicillin. Plasmids were purified and recombinant integrity of inserted nucleic acids confirmed by sequencing.

3.2.7. Construction of a novel gene synthesis platform for the large scale production of small synthetic genes

An integrated gene synthesis platform was developed for the efficient production of small synthetic genes. This platform combines automation, simplicity and robustness, while decreasing the error rate associated with conventional gene synthesis methods. Initial experiments described above defined the most appropriate PCR assembly protocol. Subsequent experiments evaluated the efficacy of the protocol when applied for the simultaneously synthesis of 96 genes encoding venom peptides. Ninety-six genes were designed, using ATGenium codon optimization algorithm, by back-translating corresponding peptide sequences and by optimizing codon usage for high levels of expression in *E. coli*. Codons were selected randomly using a Monte Carlo approach according to *E. coli* codon usage of highly expressed genes. Genes were designed to have a GC content between 40% and 60% and a codon adaptation index (CAI) value higher than 0.8. The sequences of the 96 optimized genes are presented in (see Table S3. 5 in Annex). The pool of primers required to synthesise 96 synthetic genes encoding venom peptides were designed to have 50-60 nt in length, an overlap region of 20 nt between forward and reverse primers with a gap sequence of 20 nt, and included an additional 16-bp conserved sequences at 5'- terminus of both forward and reverse outer primers to allow ligation-independent cloning. The 96 genes were produced from 96 mixes of six oligonucleotides with 50-60 nt in length and 20 nt overlaps. Oligonucleotides were synthesised by Integrated DNA Technologies at the smallest scale (primer solutions at 5 µM) with desalting purification. The two outer primers were used at a final concentration of 800 nM while the inner primers were pooled together in an equimolar mixture to achieve the final concentration of 20 nM. Primer dilutions and PCR assembly was carried out in a 96-well plate format using a Tecan workstation (Switzerland). KOD Hot Start DNA polymerase (EMD Millipore) was used for PCR assembly using optimized conditions to minimize primer-dimer formation and nonspecific amplifications. PCR reactions were performed in a 50 µL total volume and consisted of 0.2 mM dNTPs, 1.5 mM $MgCl_2$, 1x reaction buffer and 1 unit of KOD Hot Start DNA polymerase (EMD Millipore). PCR assembly reactions were carried out in a 96-well PCR plate format. The cycling parameters were as follows: 1 cycle of 95°C for 2 min; 26 cycles of 95°C for 20 s, 60°C for 8 s, and 70°C for 3 s. After PCR assembly, PCR products were visualized by agarose gel electrophoresis and purified through silica-base chromatography in a Tecan liquid handler (Switzerland). Purified PCR products were cloned into pHTP1 expression vector (NZYTech, Ltd) using the NZYEasy cloning kit (NZYTech, Ltd) that follows a *ligase free* technology. Gene assembly products were mixed with 120 ng of linearized vector, using a molar ratio of 1:5 (vector:insert). Cloning reactions were performed in 10 µL volume in a 96-well PCR plate and preceded for 1 h at 37ºC on a heating block. The mixtures were then incubated at 80ºC for 10 min followed by 10 min at 30ºC. Recombinant plasmids were transformed using a high-throughput method into *E. coli*

DH5α competent cells and spread on LB agar plates supplemented with 50 µg/mL of kanamycin. After overnight incubation at 37ºC, only one colony per transformation was picked and grown in 5 mL of LB kanamycin medium in 24-deep-well plates sealed with gas-permeable adhesive seals. Cultures were incubated at 37ºC for ~16 h and cells were then harvested at 1,500 ×*g* for 15 min. Plasmids were purified from bacterial pellets in a Tecan workstation (Switzerland), and subsequently the DNA sequence of each gene was verified by Sanger sequencing. In case the DNA sequence did not correspond with the designed gene, a second and eventually third colony was picked for sequencing analysis.

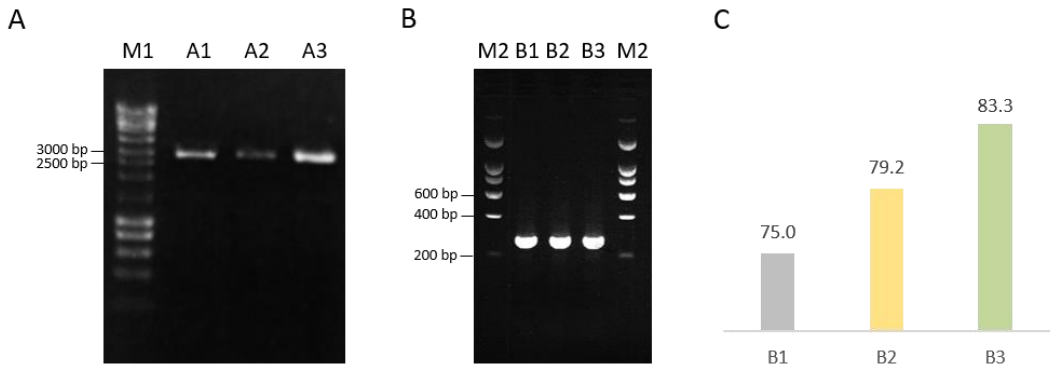
3.3. Results and discussion

3.3.1. Synthesis and assembly of the 213 nt gene (gene A) encoding *alpha-elapitoxin-Nk2a* toxin using PCA-DT and PCA-DTF methods

The gene encoding the toxic peptide *alpha-elapitoxin-Nk2a* was designed to maximize expression in *E. coli*. Initial experiments aimed identifying the most appropriate strategy for the synthesis of small genes and attempted to reduce the number of steps involved in traditional gene synthesis approaches. Thus, the efficiency of PCA-DT and PCA-DTF PCR-based methods was tested for the synthesis of gene A, which has a size of 290 nt (Figure 3.1). To ensure simplicity and speed in synthetic gene production, PCA-DT strategy does not involve an additional cloning step. Using method A1 (Figure 3.1), gene A was synthesised using pNZY28 vector as DNA template and employing two long oligonucleotides (135 nt) to amplify the full-length plasmid sequence containing half of the gene sequence in each 5' and 3' ends. Long oligonucleotides are more prone to incorporate errors. Thus, gene A was also amplified using a set of 14 and 12 overlapping oligonucleotides (Figure 3.1, strategies A2 and A3, respectively), which contain a 20 nt sequence that hybridises with cloning plasmid. In contrast, PCA-DTF methods use a template-less approach to assemble the toxic gene. The six methods described in Figure 3.1 were employed to synthesise gene A. The data, presented in Figure 3.2, A and B, revealed that all strategies effectively generated the toxic gene. However, yield of the target nucleic acid assembled through strategies B were higher when compared with the quantity of PCR product obtained using the PCA-DT. Although strategy A does not require an additional step to insert the synthetic gene into the cloning plasmid, this strategy was more tedious due to the long periods required for nucleic acid amplification that involved PCR of both the toxic gene and the vector (~3 Kb, pNZY28 vector plus target gene). In addition, cloning efficiency, evaluated through colony-PCR, revealed an approximately 95% cloning efficacy of the Gateway system versus 80% when using the self-ligation process of strategy A.

Figure 3.2| The efficacy of PCA-DT and PCA-DTF methodologies to generate small nucleic acids.

PCA-DT strategy PCA-DTF strategy



Panel A: synthesis of gene A was performed using a vector DNA template (A1, A2 and A3) and the PCR products correspond to a 3,099 bp DNA fragment, which combines gene A sequence plus pNZY28 vector sequence. B: gene A was also assembled using different pools of overlapping oligonucleotides (B1, B2 and B3). C: Effect of primer design on percentage of clones without errors. M1: NZYDNALadder III (NZYTech, Ldt), M2: NZYDNALadder I (NZYTech, Ldt).

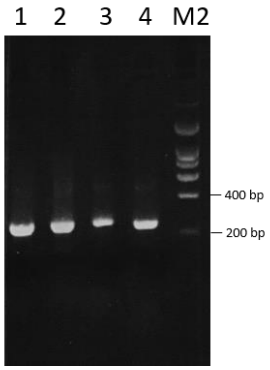
These results suggest that strategies including a DNA template and involving plasmid re-circularization are probably not the best option for the synthesis of small genes as they are more tedious while leading to lower cloning efficiencies. These two issues are particularly important when protocols require automation. In order to verify if the length of the oligonucleotides influences the appearance of errors in synthetic genes, we selected 24 clones synthesised by strategies B1, B2 and B3 for DNA sequence verification using Sanger sequencing. Approximately 80% of the sequenced clones for each one of the three strategies presented the correct DNA sequence (Figure 3.2, C). Thus, increasing primer size from 40 to 60 nt has no impact in the number of errors in the resulting synthetic DNA. Therefore, taken together the data suggest that the PCA-DTF gene assembly method that uses a set of 60 nt overlapping oligonucleotides with 20 nt gaps, is the most convenient strategy for the synthesis of small genes as it provides high gene synthesis and cloning efficiencies.

3.3.2. Performance of various thermostable DNA polymerases for gene synthesis

Taq, *Kod* and *Pfu* polymerases have been commonly used for the production of synthetic genes following PCR-based methods (Stemmer *et al*., 1995; Xiong *et al*., 2004; Young & Dong, 2004). However, *Taq* polymerase is known to be error-prone (Tindall & Kunkel, 1988). The use of *Kod* and *Pfu* polymerases allow more accuracy in gene synthesis methods although the elongation rate of *Pfu* polymerases is lower (Takagi *et al*., 1997; G. Wu *et al*., 2006). Here, the efficacy of four different DNA polymerases (KOD Hot Start DNA polymerase, Q5® Hot Start High-Fidelity DNA Polymerase, Pfu Turbo DNA polymerase, and Taq DNA polymerase) for the production of the gene B (260 bp), using B3 strategy PCA-DTF (described above) was

analysed. The data, presented in Figure 3.3, revealed that the four polymerases effectively assembled the 260 bp gene. However, KOD Hot Start DNA polymerase seems to express a higher performance when compared with the other enzymes. These data suggest that efficacy of *Kod* DNA polymerases is higher than *Taq* and *Pfu* enzymes for assembling small genes. After cleaning the PCR products, genes were cloned into pDONR201 vector and 24 clones assembled by each one of the four DNA polymerase were sequenced.

Figure 3.3| Performance of four thermostable DNA polymerases for the synthesis of gene B using PCA-DTF.



Lane 1: KOD Hot Start DNA polymerase; lane 2: Q5® Hot Start High-Fidelity DNA polymerase; lane 3: Pfu Turbo DNA polymerase and lane 4: Taq DNA polymerase. M2: NZYDNALadder I (NZYTech, Ldt).

The data, presented in Table 3.2, revealed that appearance of mutations is more frequent for *Taq*, followed by Q5® Hot Start HF and *Pfu* Turbo DNA polymerase (Table 3.2). In contrast, only five out of the 22 recombinant plasmids containing the synthetic gene B assembled by KOD Hot Start DNA polymerase presented errors, which reflects one mutation per 1.15 kb. Deletions and substitutions were the most frequent errors identified in the 96 variants of gene B sequenced. As expected, the data suggested that KOD Hot Start DNA polymerase is more accurate than the other three DNA polymerases due to a higher fidelity. Moreover, the PCA-DTF procedure appears to be very efficient with KOD Hot Start DNA polymerase due to the high elongation rate of this DNA polymerase; completion of the gene synthesis protocol is achieved in less than 40 minutes.

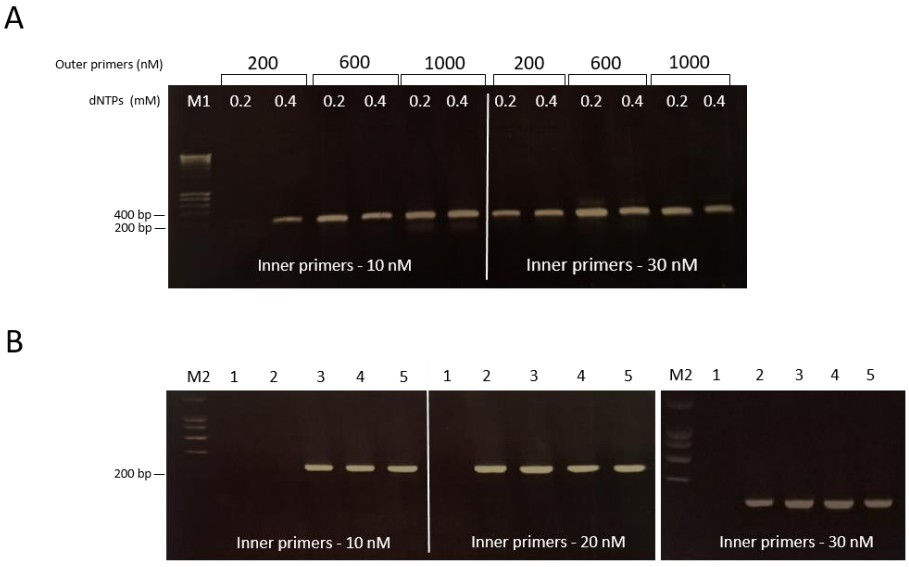
Table 3.2| Gene synthesis error rates when were used different DNA polymerases.

DNA polymerase	Hot Start activity	Number of clones sequenced	Bases sequenced	Number of bases deleted, inserted or substituted	Error rate (error/kb)	Number of deletions	Number of insertions	Number of substitutions
KOD Hot Start DNA pol.	yes	24	5720	5	0.87	2	1	2
Q5® Hot Start High-Fidelity DNA pol.	yes	24	6240	10	1.60	6	1	3
Pfu Turbo DNA pol.	no	24	6240	10	1.60	5	2	3
Taq DNA pol.	no	24	5980	13	2.17	2	3	7
Total		96	24,180	38		15	7	15

3.3.3. Oligonucleotide concentration influences the efficacy of gene synthesis

Since PCA-DTF is suggested to be the best method to produce small synthetic genes and KOD Hot Start DNA polymerase the most effective enzyme, we analysed the influence of oligonucleotide concentration on gene assembly efficiency. Thus, three different concentrations of inner oligonucleotides were combined with different concentrations of outer primers in a PCR-assembly reaction set to synthesise gene B. Initially, concentrations of inner oligonucleotides were of 10 nM and 30 nM, and outer oligonucleotides were tested at 200, 600 and 1000 nM. After the assembly of gene B the resulting nucleic acids were analysed by agarose gel electrophoresis. The data, presented in Figure 3.4, A, suggest that gene synthesis is most effective with 30 nM of inner primers. In addition, the best concentrations of outer primers were of 600 and 1000 nM. In order to define more precisely the best concentrations of primers, outer primer concentration was fixed at 800 nM and concentrations of inner primers varied from 10 to 30 nM. Data suggest that at 800 nM of outer primers, the optimal concentration of inner oligonucleotides is 20 nM (Figure 3.4, B). The assembly reaction was also performed under different concentrations of dNTPs. Interestingly, the results suggest that concentrations of dNTPs below 0.2 mM are not appropriate for gene synthesis. Thus, for a robust and successful PCA-DTF procedure a concentration of 0.2-0.4 mM of dNTPs seems to be the most effective.

Figure 3.4| Influence of oligonucleotide concentration in the efficacy of the assembly reaction.

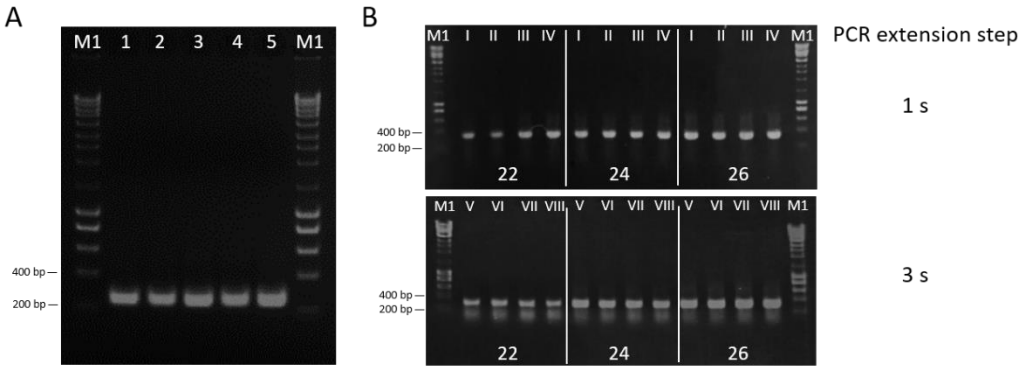


A: PCR assembly was performed with 200, 600 and 1000 nM of outer oligonucleotides using 0.2 or 0.4 mM dNTPs and 10 nM or 30 nM of inner primers. B: Outer primers concentration was fixed to 800 nM and inner primer concentration varied between 10 to 30 nM. Lanes 1-5 correspond to: 0.1, 0.2, 0.3, 0.4 and 0.5 mM dNTPs. M1: NZYDNALadder III (NZYTech, Ldt); M2: NZYDNALadder I (NZYTech, Ldt).

3.3.4. Effect of cycling temperatures on the efficiency of gene synthesis

In the previous assembly reactions following PCA-DTF, annealing temperature was set to 60ºC. Several studies (Hoover & Lubkowski, 2002) have suggested that factors such as melting temperature (T_m) and GC content affect optimal assembly. Thus, the efficiency of synthesis of gene B was tested using a gradient of annealing temperatures (50ºC, 52.3ºC, 54.6ºC, 59.6ºC and 62ºC) applying the optimal PCR conditions described in the previous section. The data, presented in Figure 3.5, A, suggest that oligonucleotide assembly occurs at temperatures ranging from 50 to 62ºC, although yields of nucleic acid seem to increase at higher annealing temperatures. In addition, the effect of the number of PCR cycles in the efficiency of gene synthesis was tested by synthesizing gene C using 22, 24 and 26 thermal cycles.

Figure 3.5| Effect of cycling temperatures on the efficiency of gene synthesis.



Panel A: different annealing temperatures were studied (1: 50ºC; 2: 52.3ºC; 3: 54.6ºC; 4: 59.6ºC and 5: 62ºC). Panel B: 22, 24 and 26 cycles combined with different times of denaturation, annealing and extension were used to synthesise gene C (I: 95ºC for 20 s, 60ºC for 10 s, 70ºC for 1 s; II: 95ºC for 16 s, 60ºC for 10 s, 70ºC for 1 s; III: 95ºC for 20 s, 60ºC for 8 s, 70ºC for 1 s; IV: 95ºC for 16 s, 60ºC for 8 s, 70ºC for 1 s; V: 95ºC for 20 s, 60ºC for 10 s, 70ºC for 3 s; VI: 95ºC for 16 s, 60ºC for 10 s, 70ºC for 3 s; VII: 95ºC for 20 s, 60ºC for 8 s, 70ºC for 3 s; VIII: 95ºC for 16 s, 60ºC for 8 s, 70ºC for 3 s). The extension times used for each number of cycles were 1 and 3 seconds.

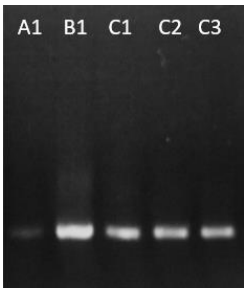
The data revealed that, as expected, the quantity of the amplified product increases with the number of thermal cycles employed (Figure 3.5, B). Likewise, when denaturation lasts for 20 s an extension of 3 s produces higher yields than when extension is performed for only 1 second. In addition, annealing of overlapping oligonucleotides during 8 seconds is more favourable than for 10 seconds. Therefore, 26 thermal cycles of denaturation (20 s), annealing (8 s) and extension (3 s) steps are optimal for PCR assembly following the PCA-DTF method.

3.3.5. Effect of oligonucleotide source in the efficacy of gene synthesis

It is well known that efficacy of gene synthesis directly depends on the quality of the synthetic oligonucleotides (Tian *et al*., 2009). Current chemical synthesis methods usually produce oligonucleotides that are prematurely terminated or comprise internal insertions or deletions

(Tian *et al*., 2009). To determine the effects of the oligonucleotide source on the production of error-free DNA fragments, gene C was synthesised using desalted, reverse-phase cartridge and reverse-phase HPLC purifications, obtained from three different suppliers. Gene C was assembled using PCR conditions defined above and five different oligonucleotide sources were analysed. The results, presented in Figure 3.6, show that oligonucleotides from supplier B displayed the best performance.

Figure 3.6| Assembly of gene C using oligonucleotides obtained from three different suppliers (A, B and C) and three purification methods.



Numbers correspond to desalted (1), reverse-phase cartridge (2) and HPLC (3) primers purification methods.

DNA plasmids of 16 recombinant clones for each condition were analysed by Sanger sequencing. The highest percentage of clones without errors was identified in genes synthesised with primers from supplier B, which were not subjected to any purification (Table 3.3). PCR products assembled using reverse-phase cartridge and HPLC oligonucleotides have a lower percentage of clones without errors (B2 - 50% and B3 - 56%, respectively) when compared with exclusively desalted oligonucleotides. Thus, it is noteworthy that oligonucleotide purification does not solve the mutation problem. The most frequent mutation identified in the 80 recombinant clones was a single base deletion (44%, see Table 3.3). These results suggest that the truncated versions of oligonucleotides (*n-1*) are difficultly removed by purification methods whereby the desalted oligonucleotides from supplier B may contain a lower frequency of deletions.

Table 3.3| Oligonucleotides purification and source used to assemble gene C.

Oligo source	Supplier	Purification method	Number of clones sequenced	Clones without errors (%)	Number of bases deleted, inserted or substituted	Number of deletions	Number of insertions	Number of substitutions
A1	A	desalted	16	68	6	4	1	1
B1	B	desalted	16	80	3	2	0	1
C1	C	desalted	16	71	7	3	2	2
C2	C	cartridge	16	50	14	4	5	5
C3	C	HPLC	16	56	11	5	4	2
Total			80		41	18 (18/41=44%)	12 (12/41=29%)	11 (11/41=27%)

3.3.6. Large scale synthesis of genes encoding venom peptides using an automated platform

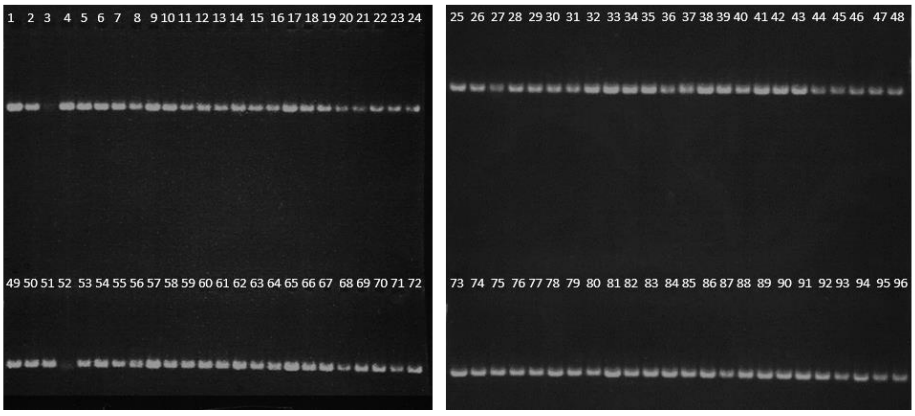
Previous optimized protocols were used to develop a platform for the simultaneous synthesis of small genes. Thus, the primary sequence of 96 venom peptides was used to design 96 genes that contained an average GC content of 49% and an average CAI of 0.86 (Table 3.4). To assemble the 96 genes, 576 (6 primers x 96 genes) oligonucleotides with a maximum of 60 bp were designed with an overlap region of 20 bp and a gap of 20 bp. In average, the genes had 240 bp in length and oligonucleotides were acquired without additional purification.

Table 3.4| Properties of 96 optimized genes that were synthesised using the HTP gene synthesis platform.

	Length (nt)	GC content (%)	Codon adaptation index (CAI)
Mean (±SD)	240 ± 9	48.9 ± 3.4	0.86 ± 0.03
Maximum	254	57	0.97
Minimum	215	40	0.80

Each gene was PCR assembled using the KOD Hot Start DNA polymerase. Outer primers were used at 800 nM (forward and reverse) while inner primers at a final concentration of 20 nM. PCR assembly was performed in 26 cycles of 95ºC for 20 s, 60ºC for 8 s and 3 s at 70ºC. The 96 genes were assembled simultaneously in a 96-well PCR plate and resulting nucleic acids analysed through agarose gel electrophoresis. The data, presented in Figure 3.7, revealed that 94 out of the 96 genes were effectively assembled representing a 98% success rate of the gene synthesis protocol when applied to a large scale.

Figure 3.7| Agarose gel electrophoresis of 96 nucleic acids encoding venom peptides assembled simultaneously using a large scale gene synthesis platform.



98% of gene products presented high yield and correct size. 2 out the 96 genes (3 and 52) showed PCR fragments with low yield. However, the 96 PCR products were used for subsequent cloning reaction and resulting plasmids were employed to transform *E. coli* cells. Resulting recombinant plasmids were purified and the integrity of DNA sequences was verified by Sanger sequencing.

After purification of the generated 96 PCR products, individual genes were subcloned into pHTP1 expression vector using a LIC method. The robustness and effectiveness of the pipeline was demonstrated when recombinant plasmids were sequenced to verify gene integrity. The initial screen of one clone per gene revealed that 77 genes (80%) were correct (Table 3.5). For 17 genes (18%) two clones were screened to identify an error-free DNA fragment. Finally, for 2 genes (2%) it was necessary to pick a third clone to obtain a correct DNA sequence. Thus, even for the two genes that apparently were amplified at a lower concentration it was possible to obtain a correct clone. In total, 26 mutations were identified in incorrect genes, leading to an overall error rate of 1 mutation per 0.9 kb. The majority of the identified mutations were deletions (77%), as it is expected from the incorporation of prematurely terminated oligonucleotide (LeProust *et al*., 2010). The remaining mutations were single-base substitutions (19%) and insertions (4%).

Table 3.5| Properties of primers used in gene synthesis of 96 genes encoding venom peptides and error rate determined for the integrated HTP gene synthesis platform developed in this study.

	HTP gene synthesis of 96 genes
Number of genes	96
Number of primers used in PCR-assembly	6
Primer length (nt)	57-60
Primer overlap/gap (nt)	20
Bases sequenced	23,058
Number of bases deleted, inserted or substituted	26
Error rate (error/kb)	1.13
Number of deletions (N; %)	20; 77%
Number of insertions (N; %)	1; 4%
Number of substitutions (N; %)	5; 19%
Genes with 1 clone sequenced (N; %)	77; 80%
Genes with 2 clones sequenced (N; %)	17; 18%
Genes with 3 clones sequenced (N; %)	2; 2%

3.4. Conclusions

The ability to *de novo* synthesise DNA sequences is rapidly emerging to improve the speed, accuracy and simplicity of recombinant DNA technology. Here, we have optimized a novel gene synthesis large scale platform for the efficient production of small genes (< 0.5 kb). The genes were directly cloned into an *E. coli* expression vector using a completely automated protocol. This gene synthesis approach presents high efficiencies of PCR assembly and cloning while revealing low error rates. The error rate of the large scale method described here is of 1.1 mutations per kb. Low error rates avoid additional steps for the removal of errors from

synthesised genes, such as the enzymatic removal of DNA mutations that uses proteins involved in the recognition of mismatches within DNA sequences. The identification of 100% correct genes was performed by screening a maximum of 3 colonies. Thus, the labour required for the selection and validation of recombinant clones is reduced. The use of overlapping oligonucleotides combined with *Kod* DNA polymerase provides a powerful alternative to conventional synthesis protocols. The length of all oligonucleotides is below 60 bp, with 20-bp overlap regions and gaps of 20 bp. This represents a decrease in the number of oligonucleotides required to synthesise a given gene, saving costs. The PCR-based gene synthesis method described here is an optimization of the simplified gene synthesis method (SGS) (G. Wu *et al*., 2006). However, among other details, the primer length used of the protocol described in this study is larger than in SGS method. In conclusion, the gene synthesis approach described here is a simple, accurate and robust system that can be used to construct, at low cost and in short periods, large numbers of *de novo* DNA molecules for a variety of applications.

4. T7 ENDONUCLEASE I MEDIATES ERROR CORRECTION IN ARTIFICIAL GENE SYNTHESIS

Ana Filipa Sequeira[1,2], Joana L.A. Brás[2], Catarina I.P.D. Guerreiro[2], Renaud Vincentelli[3] and Carlos M.G.A. Fontes[1,2]

[1] Centro Interdisciplinar de Investigação em Sanidade Animal (CIISA) - Faculdade de Medicina Veterinária, Universidade de Lisboa, Avenida da Universidade Técnica, 1300-477 Lisboa, Portugal; [2] NZYTech Genes & Enzymes, Campus do Lumiar, Estrada do Paço do Lumiar, Edifício E, r/c, 1649-038 Lisboa, Portugal; [3] Unité Mixte de Recherche (UMR) 7257, Centre National de la Recherche Scientifique (CNRS) Aix-Marseille Université, Architecture et Fonction des Macromolécules Biologiques (AFMB), Campus de Luminy, 163 Avenue de Luminy, 13288 Marseille CEDEX 09, France.

Adapted from: Sequeira *et al*. (2016) Molecular Biotechnology, 58 (8-9), 573-84.

Abstract

Efficacy of *de novo* gene synthesis largely depends on the quality of overlapping oligonucleotides used as template for PCR assembly. The error rate associated with current gene synthesis protocols limits the efficient and accurate production of synthetic genes, both in the small and large scales. Here, we analysed the ability of different endonuclease enzymes, which specifically recognize and cleave DNA mismatches resulting from incorrect impairments between DNA strands, to remove mutations accumulated in synthetic genes. The *gfp* gene, which encodes the Green Fluorescent Protein (GFP), was artificially synthesised using an integrated protocol including an enzymatic mismatch cleavage step (EMC) following gene assembly. Functional and sequence analysis of resulting artificial genes revealed that number of deletions, insertions and substitutions was strongly reduced when T7 endonuclease I was used for mutation removal. This method diminished mutation frequency by 8-fold relative to gene synthesis not incorporating an error correction step. Overall, EMC using T7 endonuclease I improved the population of error-free synthetic genes, resulting in an error frequency of 0.43 errors per 1 kb. Taken together data presented here reveal that incorporation of a mutation removal step including T7 endonuclease I can effectively improve the fidelity of artificial gene synthesis.

4.1. Introduction

The *de novo* assemblage of DNA molecules is rapidly emerging as a highly powerful molecular tool to generate any desired gene sequence (Chao, Yuan, & Zhao, 2014). Artificial gene synthesis technologies do not require preexisting DNA templates which is becoming pivotal to explore accumulating genomic and metagenomic information for which natural sequences are

difficult to access. Gene synthesis is also changing established paradigms within the recombinant DNA technology field, in particular for heterologous gene expression, vaccine development, gene therapy and molecular engineering. In recent years, improvements in artificial DNA production methodologies have originated more robust, simple, and cost-effective gene assembly technologies (Currin *et al*., 2014; Zampini *et al*., 2015). In addition, exploitation of the latest high-throughput molecular technologies supported the large-scale and low-cost production of DNA sequences (Kosuri & Church, 2014). However, current methods for *de novo* gene synthesis still display significant limitations. Prevailing DNA synthesis methodologies are based on the enzymatic assembly of chemically synthesised overlapping oligonucleotides, which span the entire length of the desired sequence. Products of the assembly reaction often contain mutations that primarily result from errors accumulated during the chemical synthesis of oligonucleotides (Ma *et al*., 2012). Thus, the major drawback to high-fidelity gene synthesis remains the quality of the single-stranded DNA oligonucleotides used for gene construction (Tian *et al*., 2009).

Various strategies have been developed to reduce the number of errors observed in synthetic genes. Despite recent progresses achieved in the production of high quality oligonucleotides, error removal from synthetic genes during or after gene assembly remains highly required. Current methods used for the correction of DNA sequences rely on the detection and correction of mismatches present in hetero-duplex DNA molecules using mismatch-binding proteins, mismatch cleavage by endonucleases or site-directed mutagenesis (Ma *et al*., 2012; Tian *et al*., 2009). Although only mutation correction with site directed mutagenesis offers total fidelity for gene synthesis, approaches incorporating mismatch active enzymes are less laborious, cost-effective and easily adapted to the large scale. Hence, there is a considerable degree of evidence suggesting that enzymatic mismatch cleavage (EMC) using endonuclease enzymes is the most promising approach to effectively reduce the levels of inaccuracies within synthetic nucleic acids. Conceptually this method is based on mismatch cleavage through the action of specific endonucleases, in particular those that recognize and cleave DNA at mismatches in hetero-duplex molecules, followed by a second DNA fragment assembly step. In combination with DNA polymerases presenting 3'-5' exonuclease activity, also termed proofreading DNA polymerases, this approach was shown to be very effective (Fuhrmann *et al*., 2005). However, it remains largely unknown which class of mismatch specific endonucleases are most effective for error removal during gene synthesis.

The family mismatch-specific nucleases includes: 1) single-strand specific nucleases, such as S1 and P1 nucleases, mung bean nuclease, and CEL I nuclease (Desai & Shankar, 2003; B. Yang *et al*., 2000); 2) mismatch repair endonucleases, such as MutH (Smith & Modrich, 1997), and 3) resolvases, such as phage T4 endonuclease VII, T7 endonuclease I and *Escherichia coli* endonuclease V (Ma *et al*., 2012). All these enzymes have a specific activity towards DNA molecules (Fuhrmann *et al*., 2005; Saaem *et al*., 2012; Till, Burtner, Comai, & Henikoff, 2004).

T7 endonuclease I is a bacteriophage resolvase that has been extensively used in the detection of single-base pair mismatches and mutational screening. In addition, this endonuclease was suggested to recognize all types of mismatches, including those occurring in small hetero-duplex loops (Babon, McKenzie, & Cotton, 2003; Tsuji & Niida, 2008). However, Fuhrmann and colleagues (2005) have reported that T7 endonuclease I failed to cleave selected mispairs. Thus, the specific cleavage activity of mismatch specific nucleases, in general, and of T7 endonuclease I, in particular, is poorly understood. Here, an integrated gene synthesis protocol was used to synthesise the *gfp* gene, which encodes green fluorescent protein. Different endonucleases were used in an EMC assay to reduce error rates and improve gene synthesis fidelity. The data revealed that T7 endonuclease I is highly effective to remove mutations which accumulate during artificial gene synthesis.

4.2. Materials and Methods

4.2.1. Synthesis, cloning, expression and purification of mismatch cleavage nucleases from different sources

Mismatch-specific nucleases have the ability to cleave single base pair mismatches in hetero-duplex DNA templates. In this study, six endonucleases from different sources (Table 4.1) were selected to understand the influence of enzymatic mismatch cleavage as an error correction tool during *in vitro* gene synthesis. Four of the six endonucleases chosen belong to the P1/S1 nuclease family and display strong primary sequence conservation especially at the active site, where critical catalytic residues are identical between these enzymes. The other two nucleases selected, Endonuclease V from *Escherichia coli* and T7 endonuclease I from bacteriophage T7, were reported as mismatch cleavage enzymes in different studies involving either error removal or mutation detection (Fuhrmann *et al*., 2005; Huang, Cheong, Lim, & Li, 2012). Genes encoding the six endonucleases were synthesised *in vitro*, with a codon usage optimized for expression in *Escherichia coli*, using the gene synthesis method described in Chapter 3, and cloned into pHTP1 (6HIS tag) expression vector. DNA sequences of the six *de novo* designed genes are reported in supplementary Table S4. 1. The gene encoding the maltose binding protein was also cloned in fusion with T7 endonuclease I to promote expression solubility and protein stability. The seven recombinant plasmids (T7 endonuclease I gene was present in the fused and the unfused form) were used to transform *E. coli* BL21(DE3) cells. Expression of each one of the 7 recombinant endonucleases was achieved by adding isopropyl β-D-thio-galactopyranoside (IPTG) (1 mM final concentration) to mid-exponential phase cultures and incubation for 16 hours at 16 ºC. The His_6-tagged recombinant proteins were purified from cell-free extracts by immobilized metal ion affinity chromatography (IMAC) using standard methodologies (Cheung, Wong, & Ng, 2012). Fractions containing purified proteins were analysed through SDS-PAGE.

Table 4.1| Six endonucleases selected from different sources were used for error removal in gene synthesis protocols.

Mismatch cleavage endonuclease	Enzyme family	Origin	Organism
Endonuclease V	Endonuclease	Bacteria	*Escherichia coli*
Endonuclease III-wt	S1/P1 nuclease	Eubateria	eubacterium SCB49
Endonuclease III-mut	S1/P1 nuclease	Eubacteria	eubacterium SCB49
Endonuclease I	S1/P1 nuclease	Plant	*Apium graveolens*
Endonuclease II	S1/P1 nuclease	Fungi	*Tulasnella calospora*
T7 Endonuclease I	Resolvase	Enterobacteria	Bacteriophage T7

4.2.2. Design of a 967 nt *lac-gfp* gene using overlapping oligonucleotides

An artificial gene with 967 nt was designed by combining the coding sequence of the green fluorescence protein (GFP) with the lacZ promoter (see Table S4.2 in Annex). Additional cloning sequences (16-bp sequence at 5' and 3' ends) were included in the artificial gene to facilitate ligation independent cloning (LIC). The DNA sequence encoding GFP protein was designed to display a codon usage optimized for high expression levels in *E. coli*. The 967 DNA sequence was parsed into 24 oligonucleotides of 60 nt, including 20 nt overlap regions between complementary pairs and allowing gaps of 20 nt. Oligonucleotides were synthesised by Integrated DNA Technologies (IDT) using the smallest scale available, with no purification. The sequence of the gene construct and all oligonucleotides used for the gene assembling protocol are presented in supplementary Table S4.2 and Table S4.3, respectively.

4.2.3. PCR assembly to produce synthetic nucleic acids

Synthetic genes were produced by assembly PCR. Internal oligonucleotides used in the assembling PCR reaction were grouped into a pool, termed the inner oligonucleotide mixture, and diluted to 125 nM stock solution. The two outer (or external) forward and reverse primers were used at higher final concentrations (800 nM). The first PCR (PCR1) was performed in a final volume of 50 µL using 1 unit of KOD Hot Start DNA polymerase (EMD-Millipore), 1x reaction buffer provided by the enzyme manufacturer, 0.2 mM dNTPs and 1.5 mM $MgCl_2$. Outer and inner oligonucleotides were used at final concentrations of 800 and 20 nM, respectively. PCR1 cycling parameters were 95ºC for 2 min, followed by 15 cycles of denaturation at 95ºC for 20 s, annealing at 55ºC for 10 s and extension at 70ºC for 15 s. A 5 µL aliquot of resulting PCR1 product was used as template to perform a second PCR (PCR2). PCR2 was performed incorporating exclusively outer oligonucleotides to ensure the production of full-length variants of the gene of interest. PCR2 was carried out in a final volume of 50 µL containing 1 unit of KOD Hot Start DNA polymerase (EMD-Millipore), 0.2 mM dNTPs, 1.5 mM $MgCl_2$ and 250 nM of each outer primer. The PCR conditions were 1 cycle at 95ºC for 2 min,

30 cycles at 95ºC for 20 s, 60ºC for 10 s and 70ºC for 20 s. Amplified nucleic acids from PCR2 were visualized by agarose gel electrophoresis and purified using silica-based columns.

4.2.4. Endonuclease activity assay

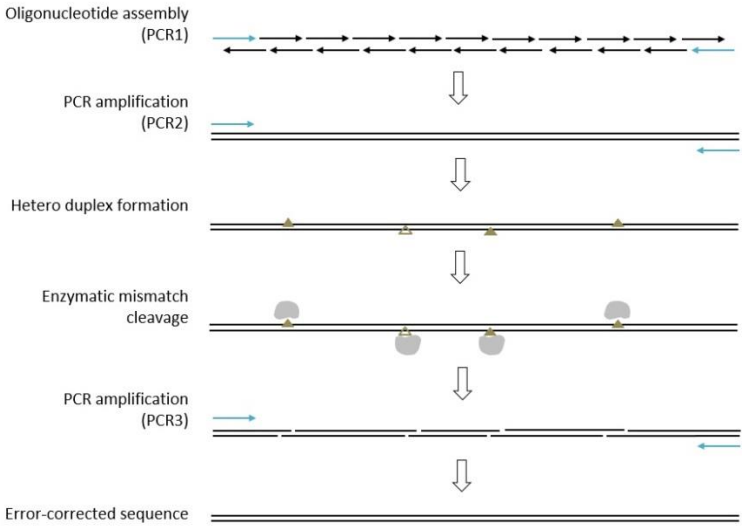
The 967 nt synthetic PCR product obtained as described above was used to analyse the cleavage activity of the seven recombinant endonucleases. Enzymatic activity was tested by incubating 25 ng of the nucleic acid in a standard reaction buffer (50 mM NaCl, 10 mM Tris-HCl, 10 mM $MgCl_2$, 1 mM $ZnCl_2$, 1 mM DTT, pH 7.9) and 1 µL (5 pmol) of each recombinant endonuclease, at 25ºC for 1 hour. The endonuclease activity of three commercial enzymes was also accessed following the supplier's recommendations. The enzymes used as controls were S1 nuclease (ThermoFisher Scientific), T7 endonuclease I (New England Biolabs) and CorrectASE (Invitrogen) and were incorporated in the reaction at a 1 µL volume. After incubation/digestion, reaction products were heated for 20 min at 65ºC to inactivate nuclease activity and resulting nucleic acids integrity analysed through agarose gel electrophoresis (1.5% w/v). The efficacy of the different enzymes to degrade 50 ng of plasmid DNA (pNZY28) in quantities ranging from 13.5 to 0.5 pmol was subsequently evaluated.

4.2.5. Error removal by enzymatic cleavage of DNA mismatches

An enzyme treatment step involving the use of mismatch cleavage nucleases was designed to increase the percentage of error-free DNA fragments resulting from PCR assembly reactions. Thus, the products resulting from PCR2 were used in an enzymatic mismatch cleavage (EMC) assay. Resulting nucleic acids were employed as template in a final PCR reaction (PCR3) to ensure that only DNA fragments with correct sequences were amplified. The complete workflow of the protocol employed for gene synthesis and enzymatic error removal is presented in Figure 4.1. To produce incorrect impairment between DNA bases, also termed DNA mismatches, which act as substrates for mismatch endonuclease enzymes, PCR2 products were denatured and re-annealed (Figure 4.1). Thus, PCR2 products were diluted to 25 ng/µL in standard reaction buffer (50 mM NaCl, 10 mM Tris-HCl, 10 mM $MgCl_2$, 1 mM $ZnCl_2$,1 mM DTT, pH 7.9) and DNA was denatured at 98ºC for 2 min and slowly re-hybridized by reducing the temperature down to 4ºC in order to obtain hetero-duplex nucleic acids. Samples remained at 4ºC for 5 min, followed by 5 min at 37ºC and a final step at 4ºC. To proceed with the EMC reaction, 10 µL of the re-annealed DNA was incubated with 1 µL of each endonuclease studied. Precisely 13.5 or 2.7 pmol of the two T7 endonuclease I recombinant derivatives were used in the cleavage reaction. Commercial enzymes used as controls, namely T7 endonuclease I-control 1 (New England Biolabs) and CorrectASE-control 2 (Invitrogen) were used at 1 µL, thus at an unknown concentration. A negative reaction that did not incorporate nucleases but contained exclusively hybridized DNA was also incubated with 1 µL of 1x reaction buffer. Reactions were allowed to proceed for 1 hour at 25ºC. After

incubation, reactions were heated 20 min at 65ºC to inactivate the nucleases. A final PCR reaction (PCR3) was performed combining 2 µL of digestion reaction under similar conditions to PCR2 and as described above. PCR3 cycling conditions were of 1 cycle at 95ºC for 2 min and 30 cycles at 95ºC for 20 s, 60ºC for 10 s and 70ºC for 20 s. Synthesised genes were gel purified following standard protocols.

Figure 4.1| Gene synthesis workflow including an endonuclease mismatch cleavage assay.



Overlapping oligonucleotides are assembled (PCR1) and used as template for a second PCR reaction (PCR2) to build the full-length nucleic acid. A denaturation-renaturation step is used to form hetero-duplexhetero-duplexes DNA containing mismatches. DNA mismatches are recognized and cleaved by endonucleases. Using the digestion reaction as template a third PCR reaction is used (PCR3) to recover the error-corrected DNA fragment.

4.2.6. Functional analysis and sequencing of synthetic *gfp* gene

The error-removal efficacy of recombinant endonucleases during *de novo* gene synthesis of the *gfp* gene was evaluated in a functional assay and by DNA sequencing. The fluorescence of GFP protein allows a simple and expedite assessment of the success of gene synthesis, as it is likely that DNA fragments accumulating mutations will lead to the production of non-fluorescent GFP derivatives. Thus, fluorescence, visible under UV or blue light, may indicate that the resulting genes do not include mutations in the DNA sequence. After gene synthesis, full-length *gfp* gene constructs were cloned into pHTP0 cloning vector using the NZYEasy cloning kit (NZYTech, Ltd), according to the manufacturer's conditions. Recombinant plasmids were transformed into *E. coli* DH5α competent cells that were grown in LB agar plates containing 200 µg/mL ampicillin and IPTG (0.1 mM final concentration) to induce the expression of the GFP protein. After overnight incubation at 37ºC, a functional assay for the initial qualitative evaluation of the integrity of the artificial *gfp* gene sequence was performed by counting the number of fluorescent versus non-fluorescent colonies grown in the LB agar plates. In addition, 32 colonies of each endonuclease treatment were randomly selected for DNA sequencing, without regard to GFP expression. In total, 224 (32×7 treatments) plasmids

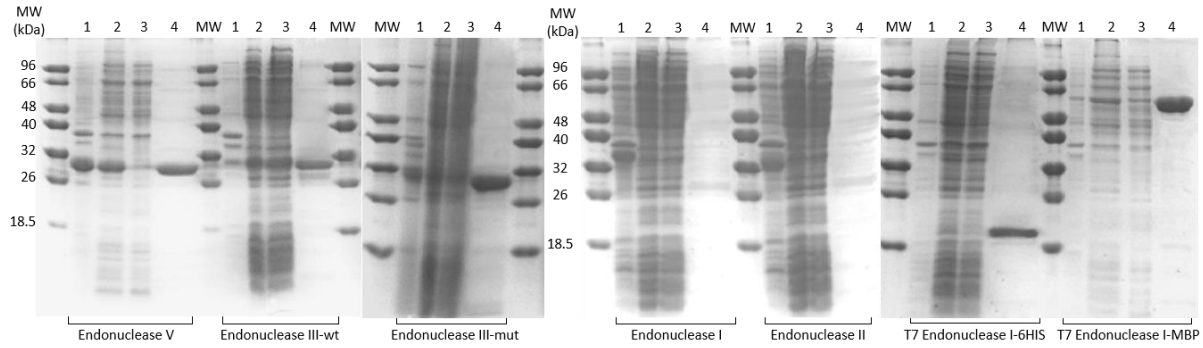
were purified from the bacterial pellet using a silica-based protocol and the integrity of resulting nucleic acids quantified by Sanger sequencing.

4.3. Results

4.3.1. Cleavage activity of recombinant endonucleases

In this study six endonucleases known to present mismatch cleavage activity were selected to identify the most efficient and accurate enzyme to incorporate in a gene synthesis protocol and contribute to error removal (Table 4.1). Some of the selected enzymes, such as T7 endonuclease I, are notorious for their capacity to cleave mismatch regions in hetero-duplex nucleic acids and have been applied in mutation screening (Huang *et al*., 2012; Till *et al*., 2004) and to a lesser extend in repair systems applied in artificial gene synthesis (Currin *et al*., 2014; Fuhrmann *et al*., 2005). The endonucleases were of prokaryotic or eukaryotic origins (Table 4.1). A mutant derivative of Endonuclease III from *Eubacterium* SCB49, which has a redesigned active site presenting conserved residues observed in Endonuclease I from plant enzymes, was also produced for these studies. Thus, the genes encoding the 6 different endonucleases were designed with a codon usage optimized for high expression in *E. coli* and artificially synthesised following developed protocol (see Chapter 3). The six artificial genes were cloned into pHTP1 expression vector. This vector incorporates a hexa-histidine (6HIS) tag to the N-terminus of the recombinant protein. Recombinant T7 endonuclease I was also expressed in fusion with an N-terminal maltose binding protein domain and an internal 6HIS tag. The seven recombinant proteins were expressed in *E. coli* and purified (Figure 4.2).

Figure 4.2| Recombinant expression and purification of DNA endonucleases in *Escherichia coli*.



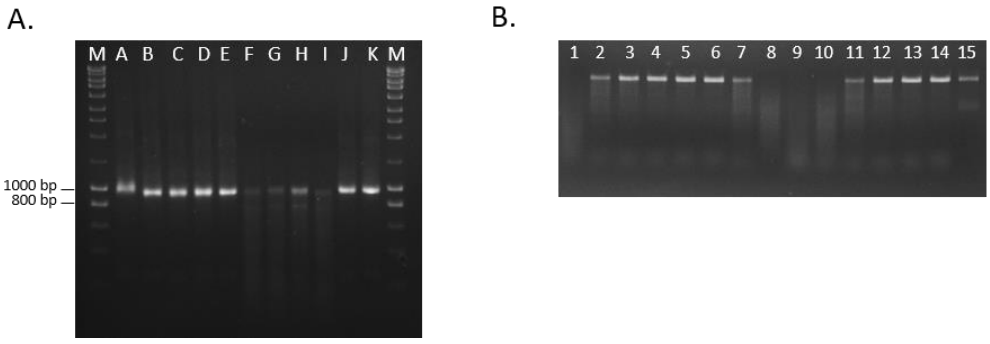
Seven endonucleases were purified through IMAC from *E. coli* cells and protein homogeneity evaluated by SDS-PAGE. Lanes 1: insoluble protein cellular extract; Lanes 2: soluble protein cellular extract; Lanes 3: protein fraction not retained by the affinity column; Lanes 4: purified recombinant endonucleases. Sizes of molecular mass protein markers are shown. The names of the seven recombinant proteins are displayed below the corresponding SDS-PAGE figure.

All five prokaryotic enzymes were expressed in the soluble form by *E. coli*, which failed to produce the two eukaryotic proteins at significant levels. These enzymes presented a

molecular mass in agreement with that deduced from primary sequence. The proteins from plant or fungal origin were found predominantly in the form of inclusion bodies. Inclusion bodies occur as a result of intracellular accumulation of partially folded expressed proteins which aggregate through non-covalent hydrophobic or ionic interactions and are commonly observed for several recombinant polypeptides expressed at high levels in *E. coli* (Rosano & Ceccarelli, 2014). Thus, considering that *E. coli* cells are unable to produce soluble forms of the two eukaryotic endonucleases analysed in this study, these enzymes were excluded from further analysis.

To determine the capacity of various endonucleases to cleave DNA molecules, the five recombinant prokaryotic proteins were incubated with a homogeneous *gfp* DNA fragment and integrity of resulting nucleic acid was evaluated through agarose gel electrophoresis. The efficacy of the recombinant nucleases was compared with those of three commercial endonucleases, T7 endonuclease I (New England Biolabs)-control 1, Correctase (LifeScience technologies)-control 2 and S1 nuclease (ThermoFisher Scientific)-control 3. Nuclease activity was analysed through the digestion of 25 ng of a 967 nt PCR product with 5 pmol of each recombinant enzyme at 25ºC for 1 h. The data, presented in Figure 4.3, A, revealed that only the T7 endonucleases I recombinant derivatives (6HIS and MBP) displayed apparent nuclease activity, presenting identical cleavage patterns when compared with control enzymes T7 endonuclease I and CorrectASE. Activity displayed by endonuclease V is difficult to interpret and suggests an increase in the size of the nucleic acid. In addition, recombinant endonuclease V, endonuclease III-wt, endonuclease III-mut, present no nuclease activity under their reaction conditions established here. These enzymes were excluded from the titration studies presented below. In order to define more precisely the optimal concentration of recombinant T7 endonuclease I required to efficiently cleave double stranded DNA fragments resulting from artificial gene synthesis, six different quantities of the two recombinant forms of the enzyme (varying from 0.5 to 13.5 pmol) were used to cleave 50 ng of plasmid DNA.

Figure 4.3| Activity of recombinant endonucleases expressed in *E. coli*.



Panel A. The capacity of different endonucleases to affect the integrity of a 967 nt PCR product was evaluated. Lane A, endonuclease V; lane B, endonuclease III-wt; lane C, endonuclease III-mut; lane F, T7 endonuclease I-6HIS; lane G, T7 endonuclease I-MBP; lane H, T7 endonuclease I-control 1; lane I, CorrectASE-control 2; lane J, S1 nuclease-control 3 and lanes D,E,K: negative control reactions, which do not incorporate endonuclease

enzymes. Lanes M: NZYDNALadder III (NZYTech, Ldt). Panel B. Effect of recombinant endonuclease concentration on enzyme activity (lanes 1-6: T7 endonuclease I-6HIS; lanes 9-14: T7 endonuclease I-MBP). Plasmid DNA (50 ng pNZY28) was digested with the two T7 endonuclease I derivatives used in different quantities (lanes 1,9 – 13.5 pmol; lanes 2,10 – 2.7 pmol; lanes 3,11 – 1.4 pmol; lanes 4,12 – 1.1 pmol; lanes 5,13 – 0.8 pmol; lanes 6,14 – 0.5 pmol) for 1 hour at 25ºC. Efficiency of recombinant endonucleases was compared with two commercial enzymes (lane 7 - T7 endonuclease I-control 1; lane 8 - CorrectASE-control 2). The negative reaction was performed without enzyme (lane 15). Lane M: NZYDNALadder III (NZYTech, Ldt).

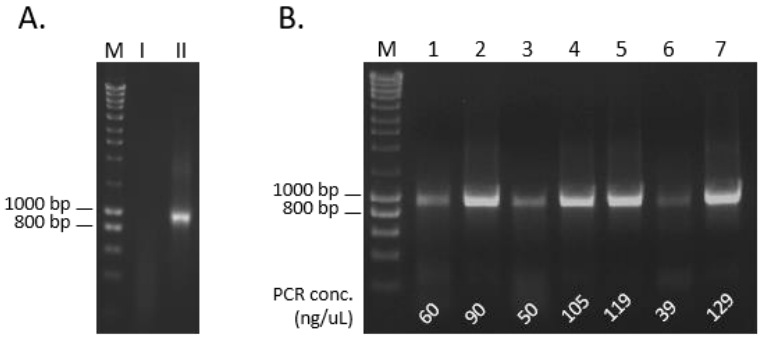
The data, presented in Figure 4.3, B, suggest that the optimal quantity of the two recombinant nucleases varied from 2.7 to 13.5 pmol. Interestingly, the nuclease activity of the two recombinant T7 endonuclease I derivatives when used at the above quantities was similar to the two commercial enzymes.

4.3.2. Error removal by enzymatic cleavage of DNA mismatches

To investigate the capacity of the two T7 endonuclease I recombinant derivatives to remove DNA mismatches arising during artificial gene synthesis, the two enzymes were incorporated in the gene synthesis protocol developed here (Figure 4.1). The protocol was used to synthesise *gfp*, which encodes a 27 kDa green fluorescent protein (GFP) that exhibits bright green fluorescence when exposed to light in the blue to ultraviolet range. The repairing step incorporated in the gene synthesis protocol occurs after PCR2 and is preceded by a denaturation/renaturation cycle required to produce hetero-duplex DNA molecules incorporating DNA-mismatches (Figure 4.1). The resulting nucleic acids were incubated with enzymes expressing nuclease activity in order to remove DNA mismatches and the integral gene recovered in a final PCR assembly phase (PCR3). The gene encoding the GFP protein was optimized for expression in *E. coli* and is controlled by an upstream lacZ promoter to ensure the expression of GFP protein in *E. coli* DH5α cells. To assemble the *gfp* gene, 24 overlapping oligonucleotides 60 nt length were designed with an overlap region of 20 nt and a gap of 20 nt and assembled through PCR. Efficacy of assembly reactions was analysed by agarose gel electrophoresis of nucleic acids resulting from PCR1 and PCR2. The data, presented in Figure 4.4, A, reveal that although no band is apparent for PCR1 a clear and specific band of the correct size is observed for PCR2. Subsequently, formation of hetero-duplex DNA was promoted through the method described by Carr (2004) by denaturation of the assembled DNA and slowly stimulating a hybridization reaction. To remove DNA mismatches from resulting nucleic acids, hybridised gene products were incubated with different mismatch cleavage endonuclease. The two recombinant T7 endonuclease I derivatives were tested at two enzyme quantities per reaction (13.5 and 2.7 pmol) and their cleavage activities compared with two control commercial proteins (T7 endonuclease I and CorrectASE). After endonuclease cleavage, reaction products were immediately used in PCR3 to recover the full length and corrected DNA sequence (Figure 4.1). The results, presented in Figure 4.4, B, revealed that when compared with the control PCR3 products generated from a

PCR reaction using a template hetero-duplex DNA not exposed to a nuclease enzyme, only three PCR3 reactions contain a significantly lower percentage of nucleic acids. These PCR reactions result from amplification of template generated after treatment with the recombinant T7 endonuclease I derivatives used at higher concentrations and the commercial enzyme CorrectASE. These results suggest that the gene products digested by these three enzymes may contain a reduced number of errors, when compared with untreated gene fragments.

Figure 4.4| Gene synthesis of *gfp* gene was performed using a set of four step reactions that include an additional error removal step.



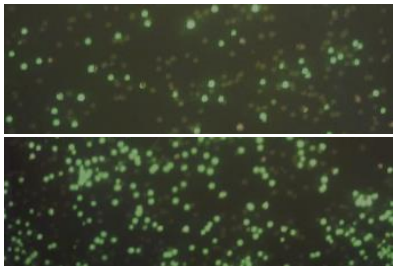
Panel A. Efficiency of oligonucleotide assembly, PCR1 (lane I), and PCR amplification, PCR2 (lane II), reactions as evaluated by agarose gel electrophoresis. Panel B. Integrity of PCR3 products obtained after mismatch cleavage and yield of each PCR product obtained. Lane 1: T7 endonuclease I-6HIS (13.5 pmol); lane 2: T7 endonuclease I-6HIS (2.7 pmol); lane 3: T7 endonuclease I-MBP (13.5 pmol); lane 4: T7 endonuclease I-MBP (2.7 pmol); lane 5: T7 endonuclease I (NEB)-control 1; lane 6: CorrectASE (Invitrogen)-control 2; and lane 7: negative reaction with no enzyme. Lanes M: NZYDNALadder III (NZYTech, Ldt).

Although gene products resulting from PCR3 after treatment with recombinant nucleases used at higher concentrations may contain fewer errors it is possible that all enzymes may have contributed to improve the fidelity of the gene synthesis process. To test this, all 7 fragments resulting from PCR3 reaction were cloned into pHTP0 and the resulting plasmids were used to transform *E. coli* cells. Fidelity of gene synthesis was initially evaluated by detecting the activity of GFP in bacterial colonies derived from the transformation reaction. Expression of GFP protein was induced by adding IPTG into LB agar plates and activity detected under blue light (Figure 4.5, A). The data, presented in Figure 4.5, A, revealed an improvement in the number of fluorescent colonies that appeared in plates generated from treated gene products when compared with plates resulting from the transformation with untreated nucleic acids. As show in Figure 4.5, B, only 31% of colonies resulting from transformation with synthetic *gfp* gene not subjected to an error removal step exhibited fluorescence. In contrast, the proportion of fluorescent colonies was increased by 2.87-fold (from 31% to 89%) when the synthetic gene was previously incubated with high concentrations of the recombinant T7 endonuclease I-MBP (Figure 4.5, B).

Figure 4.5| Analysis of GFP activity expressed by *E. coli* colonies derived from *gfp* genes artificially synthesised in the presence of different endonucleases.

A.

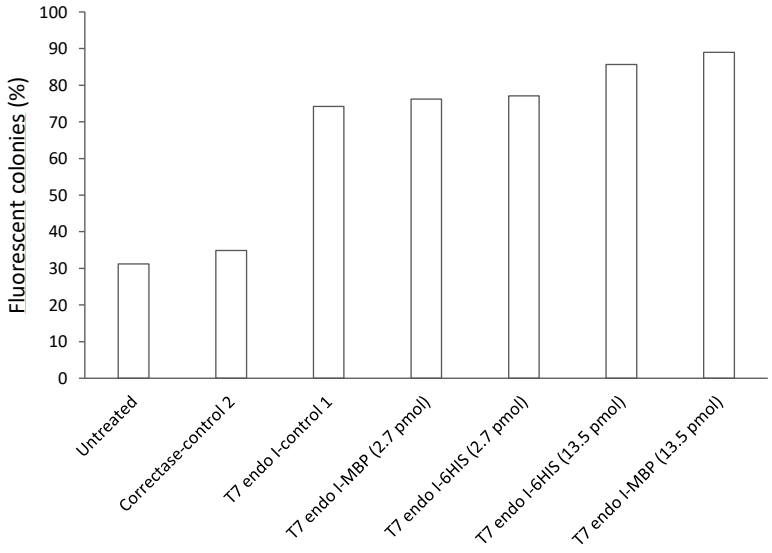
GFP expression in *E. coli* cells



Untreated DNA, 31% colonies fluorescent

DNA treated with T7 endonuclease I-MBP, 13.5 pmol, 89% colonies fluorescent

B.



Enzymatic mismatch cleavage	no	yes	yes	yes	yes	yes	yes
Fluorescent colonies (%)	31	35	74	76	77	86	89
Fluorescent colonies	594	15	1528	1516	1558	955	819
Analysed clones	1906	43	2059	1988	2021	1115	920

Panel A. Illustrative representation of GFP activity expressed by *E. coli* colonies resulting from expression of a synthetic *gfp* gene synthesised through a protocol not incorporating (with no error correction) or incorporating (with T7 endonuclease I-MBP, 13.5 pmol) an endonuclease treatment step. Colonies expressing GFP protein are green and white colonies correspond to absence of GFP expression. Panel B. The graph above represents the percentage of colonies expressing GFP activity that were transformed with genes that was subjected to a treatment with different endonucleases. The table below presents the raw data collected to produce the graph.

Percentage of fluorescent colonies generated from enzyme treated DNA was higher than the control reactions and ranged from 35 to 89%. Thus, although all enzymes seem to function effectively to remove errors accumulated during gene synthesis, the data suggest that recombinant T7 endonuclease I derivatives used at higher concentrations (13.5 pmol) constitute the most efficient enzyme treatments to reduce the percentage of mutations (Figure

4.5). Diluted versions of the recombinant T7 endonuclease I (2.7 pmol) presented an identical but slightly reduced ratio of fluorescent clones (~76%). The results suggest that detection of fluorescence constitutes a simple measurement of the success of *gfp* gene synthesis allowing to distinguish “putative error-free clones” (fluorescent colonies) from “error clones” (white colonies).

4.3.3. Error frequency of clones treated with mismatch endonucleases

Data presented above suggest that treatment with mismatch cleavage enzymes improves the fidelity of gene synthesis. However, presence of conserved mutations within *gfp* is not detected using the qualitative functional assay that evaluates GFP activity in bacterial colonies *per se*. Thus, confirmation of the integrity of the nucleic acids resulting from all enzyme treatments was further performed by sequencing. Plasmid DNA from a total of 32 colonies of each one of the seven treatments, randomly selected irrespectively of presenting GFP expression, was isolated and sequenced. Together 224 *gfp* genes were completely sequenced in both strands to identify DNA errors that surpassed mismatch cleavage treatments. The data, presented in Table 4.2 and Figure 4.6 (data from the diluted versions of T7 endonuclease I derivatives is not shown), confirmed that treatment with mismatch cleavage nucleases dramatically reduced the error frequency observed in synthetic *gfp*. Overall, the error rate, expressed by number of errors per kb of synthetic DNA, was reduced from 3.45 to 0.43. This represents a 8-fold reduction in the mutation frequency as a result of incubation with T7 endonuclease I-MBP. In general, almost all types of mutations were observed in synthetic genes, with exception of insertions with A, T and C nucleotides (Table 4.2). The type of errors identified in synthetic genes was different when the nucleic acid was exposed to different enzymes, although deletions and substitutions generally predominated (Figure 4.6). Significantly, untreated genes presented a higher frequency of deletions while this type of mutations was mostly reduced in synthetic genes exposed to the enzyme treatments. For example, error correction using T7 endonuclease I-MBP reduced the presence of single deletions (per kb) by 17 fold. When the type of deletion was analysed, the reduction was of 4.5-fold for deletion of C, 21-fold for deletion of T, 12-fold for deletion of A and no deletions of G’s were detected. Although not of this magnitude, similar levels of reduction in deletions were also verified for other enzyme treatments. To evaluate the location of the errors that survived the enzymatic mismatch cleavage, the distribution of mutations within *gfp* synthetic gene was analysed. The data, displayed in Table 4.3, suggest that the errors accumulated within untreated synthetic genes were mostly located in the core sequence. In contrast, errors that survived to error correction assays seem to be spread along the entire gene product.

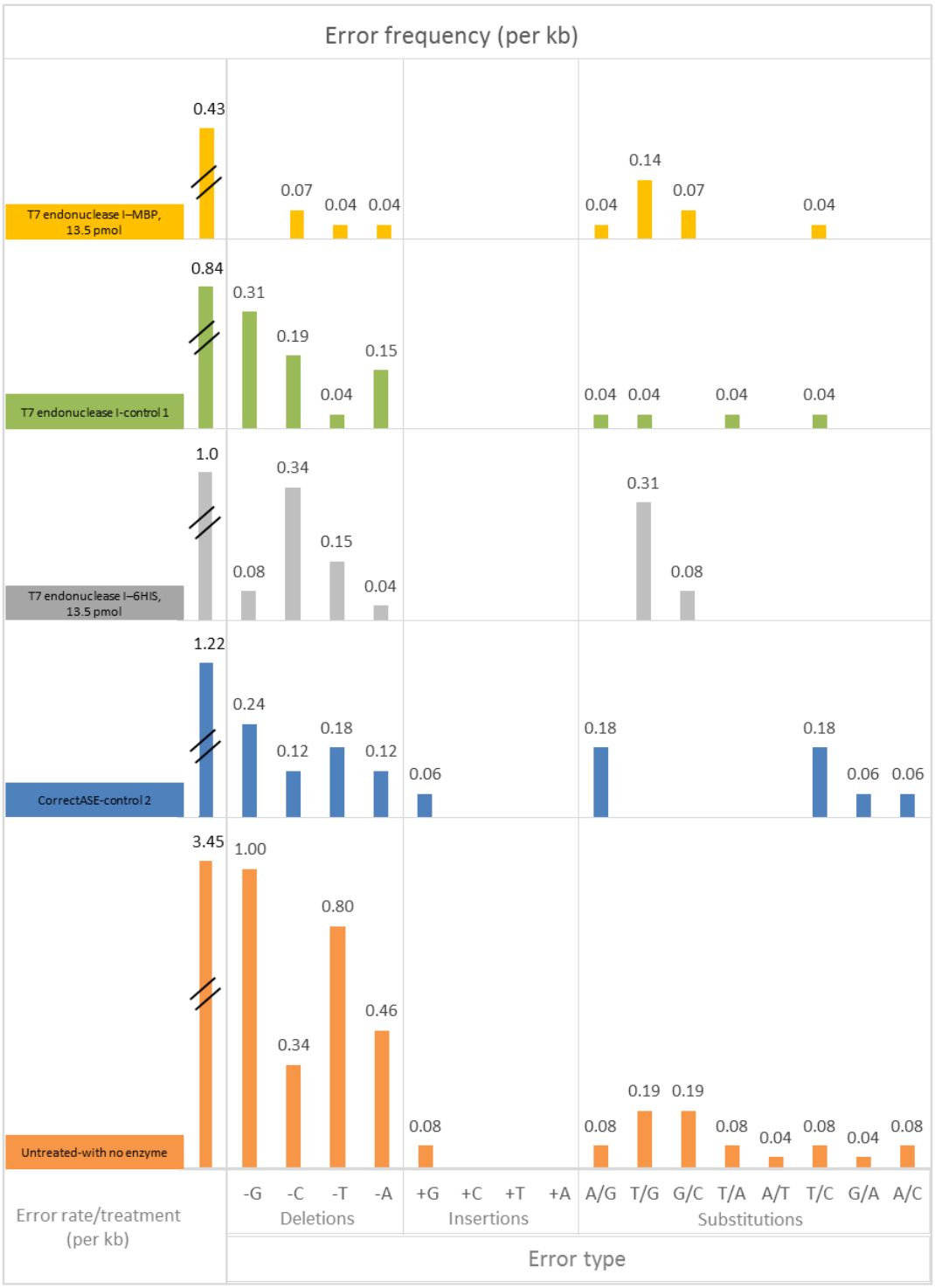
Table 4.2| Error analysis of synthetic *gfp* gene with and without error correction.

Error type	Untreated	T7 endo I (6HIS), 13.5 pmol	T7 endo I (MBP), 13.5 pmol	T7 endo I (control 1)	CorrectASE (control 2)
Deletion	68	16	4	18	11
A	12	1	1	4	2
T	21	4	1	1	3
C	9	9	2	5	2
G	26	2	0	8	4
% deletions	75.6	61.5	33.3	81.8	55
Insertion	2	0	0	0	1
A	0	0	0	0	0
T	0	0	0	0	0
C	0	0	0	0	0
G	2	0	0	0	1
% insertions	2.2	0	0	0	5
Substitution	20	10	8	4	8
Transition	5	0	0	1	8
G/T to A/T	3	0	0	1	3
A/T to G/C	2	0	0	0	0
Transversion	15	10	8	3	5
G/C to C/G	5	2	3	0	0
G/C to T/A	7	8	5	2	3
A/T to C/G	0	0	0	0	2
A/T to T/A	3	0	0	1	0
% substitutions	22.2	38.5	66.7	18.2	40
Clones sequenced	27	27	29	27	17
% clones with errors	88.9	51.9	31	55.6	70.6
Total errors	90	26	12	22	20
Bases sequenced	26109	26109	28043	26109	16439
Error frequency (per kb)	3.45	0.99	0.43	0.84	1.22

Table 4.3| Localization of errors within *gfp* synthetic gene before and after treatment with endonucleases.

Treatments	5'-end (first 60 nt)	core gene	3'-end (last 60 nt)
Untreated, with no enzyme	1	98	1
T7 endonuclease I-MBP, 13.5 pmol	8	84	8
T7 endonuclease I-6HIS, 13.5 pmol	31	65	4

Figure 4.6| Analysis of error removal efficiency of T7 endonuclease I.



Representation of error frequency per kb for untreated synthetic genes and nucleic acids artificially synthesised following a protocol incorporating an error correction step using endonucleases. The error frequency, expressed in number of mutations per kb, reflects the efficacy of each enzyme to remove errors accumulated during *de novo* gene synthesis. Error-rate for each treatment was also calculated and it is shown in the left side of the chart.

4.4. Discussion and conclusions

Gene synthesis is a powerful tool to create *de novo* DNA fragments irrespective of length and sequence. However, recurrent error rates resulting from different gene synthesis protocols have been a significant drawback to the efficient construction of DNA fragments (Kosuri & Church, 2014). Errors accumulated in synthetic genes can come from many sources. However, it is now well established that usage of imperfect synthetic oligonucleotides is the principal cause of low fidelity gene synthesis (Ma *et al*., 2012; Wan *et al*., 2014). Chemical synthesis of oligonucleotides is rarely 100% efficient (Tian *et al*., 2009). Deletions and insertions can occur in primers with a frequency of 0.5% and 0.4% per position, respectively (Tian *et al*., 2009). An improvement in the quality of oligonucleotides used for gene synthesis may result from including extra purification steps that reduce the percentage of truncated or extended molecules. However, beside the high cost of these oligonucleotides, current purifications available do not offer 100% fidelity. In addition, enzymatic gene assembly can also cause the incorporation of mutations in synthetic genes. DNA polymerases can amplify gene products with mistakes, which are replicated during gene construction. Thus, gene assembly is also error-prone although these errors are less frequent than errors resulting from incorrect oligonucleotide synthesis (Ma *et al*., 2012). Thus, there is a clear need to develop novel and effective methods to correct mutations resulting from artificial gene synthesis. Although different alternatives that minimize the number of mutations in artificial nucleic acids were previously reported, it is clear that more research is required to identify efficient enzymes to correct mutations occurring during artificial gene synthesis.

Here the efficacy of mismatch cleavage endonucleases to remove incorrect impairments of DNA strands, thus providing a mechanism to reduce mutations in artificial genes, was evaluated. Phage T7 endonuclease I was shown to present a high capacity to cleave DNA fragments. This enzyme was expressed in *E. coli* in two variants containing or not an additional MBP fusion partner. Overall data presented here revealed a higher correction activity for the T7 endonuclease I-MBP fusion protein suggesting that the 45 kDa MBP tag may improve the folding and provide further stabilization to the associated nuclease domain. In addition, T7 endonuclease I-MBP reduced by 8-fold error frequency in synthetic genes when compared with untreated samples. Deletion, insertions and substitutions were observed in all synthetic genes generated, irrespective of the enzymatic treatment. For untreated samples, the most frequent mutations observed in synthetic genes were single deletions (75.6%), followed by substitutions (22.2%) and only 2.2% of insertions. Deletions were also observed in high proportions for almost all treatments, except when T7 endonuclease I-MBP was used. Thus, this enzyme activity very effectively reduces the number of deletions in nucleic acids and in this case nucleotide substitution predominate (66.7%). Overall, the data suggest that oligonucleotide truncation as a result of inefficient chemical synthesis leads to the accumulation of a significant proportion of deletions in the artificial genes. However, it seems

that T7 endonuclease I-MBP is remarkably effective in removing single deletions in nucleic acids. Identical endonuclease activity of T7 endonuclease I was reported by Niida (2008) in studies involving mutational screening.

Although mismatch-cleavage enzymes were effective in reducing the percentage of mutations observed in synthetic genes, data presented here confirm that inclusion of an enzymatic error removal step in gene synthesis protocols is unable to completely abolish the number of errors in resulting artificial nucleic acids. Formation of hetero-duplex DNA after the PCR assembly steps of the gene synthesis protocol is key to produce the required mismatches that will be targeted by the corrective nucleases. It is clear that a fraction of DNA sequences containing mutations will re-anneal and thus will not form mismatches that result from the re-annealing with sequences containing no mutations. Eventually a cycling mismatch corrective step where hetero-duplex DNA and enzyme treatments would occur in several cycles could contribute to improve the efficacy. This would require thermal tolerant mismatch-cleavage nucleases. In addition, it is possible that complete abolition of the number of mutations observed in synthetic genes will require enzymes that are more effective. Nevertheless, data presented here suggest that the beneficial effect of adding a mismatch removal step in gene synthesis protocols is highly dependent on enzyme concentration. Lower mutation frequencies were observed when synthetic genes were treated with 13.5 pmol of T7 endonuclease I. Thus, the lower efficacy revealed by the commercial T7 endonuclease I mixture could result from an inadequate amount of enzyme used in the treatment reaction; concentration of commercial enzymes is unknown and a volume of 1 µL of enzyme was employed for mismatch-cleavage. Significantly, a higher proportion of mutations that survived enzymatic mismatch cleavage were observed at the ends of the gene. These results suggest that some of the errors that accumulate in the final gene were introduced during the final PCR amplification, namely were carried by outer primers used in PCR3. These data suggest that errors present in the outer primers used for PCR assembly will be remarkably difficult to remove from synthetic genes.

T7 endonuclease I primarily resolves four-way junctions generated by both homologous and site-specific recombination reactions by simultaneously introducing two nicks on the two non-crossing strands at 5´- sides of the junctions (Aravind, Makarova, & Koonin, 2000; de Massy, Studier, Dorgai, Appelbaum, & Weisberg, 1984). However, it is well known that this enzyme presents a broad substrate specificity which allowed its use in a variety of biotechnological applications (Aravind *et al*., 2000; White, Giraud-Panis, Pöhler, & Lilley, 1997). Data presented here revealed that fidelity of *gfp* synthetic gene was strongly improved when an additional step of enzymatic cleavage with T7 endonuclease I-MBP was integrated in the gene synthesis protocol. Error frequencies (mutations/ kb) were reduced from 3.45/kb (untreated) to 0.43/kb for samples treated with this mismatch cleavage enzyme. This improvement is related with the specific mismatch cleavage activity of T7 endonuclease I. This endonuclease most possibly recognizes the incorrect impairment of double DNA strands and cleaves DNA near to

mismatches in both strands, resulting in the creation of 5'-end in DNA fragments. The 5'-end exposed phosphate groups are important substrates for 3'-5' exonuclease digestion (Fuhrmann *et al*., 2005; Huang *et al*., 2012). The combination of T7 endonuclease I (DNA mismatch cleavage enzyme) with *Kod* DNA polymerase that has a strong 3'-5' exonuclease activity, used to generate proofreading activity, strongly contributed to reduce mutations in artificial genes. These nucleic acids were re-assembled into a full-length gene sequence through a final PCR that enriched the error-free DNA fragments in the assembly mixture.

In conclusion, inclusion of an enzymatic treatment step during the production of synthetic nucleic acids leads to a dramatically increment in the fidelity of artificial DNA sequences generated using PCR-assembly methods. Thus, the screening of integral genes from a pool of synthetic genes is facilitated by the incorporation of mismatch cleavage enzymes during gene synthesis. This approach reduces the dependence of gene synthesis fidelity on the quality of oligonucleotides used as initial templates for PCR assembly. Moreover, error removal using T7 endonuclease I derivatives leads to a more cost-effective gene synthesis and allows a simpler and quicker identification of error-free synthetic DNA products. In summary, by presenting novel evidence on the capacity of T7 endonuclease I to improve the fidelity of gene synthesis, this work opens novel avenues to explore the extraordinary potency of current gene synthesis technologies, an increasingly valuable source of nucleic acids for both fundamental and applied research.

5. GENE DESIGN, FUSION TECHNOLOGY AND TEV CLEAVAGE SITE INFLUENCE THE EXPRESSION OF DISULFIDE-RICH VENOM PEPTIDES IN *ESCHERICHIA COLI*

Ana Filipa Sequeira[1,2*], Jeremy Turchetto[3*], Natalie J. Saez[4], Fanny Peysson[3], Laurie Ramond[3], Yoan Duhoo[3], Marilyne Blémont[3], Vânia O. Fernandes[2], Luís T. Gama[1], Luís M.A. Ferreira[1,2], Catarina I.P.D. Guerreiro[2], Hervé Darbon[3], Carlos M.G.A. Fontes[1,2] and Renaud Vincentelli[3]

[1] Centro Interdisciplinar de Investigação em Sanidade Animal (CIISA) - Faculdade de Medicina Veterinária, Universidade de Lisboa, Avenida da Universidade Técnica, 1300-477 Lisboa, Portugal; [2] NZYTech Genes & Enzymes, Campus do Lumiar, Estrada do Paço do Lumiar, Edifício E, r/c, 1649-038 Lisboa, Portugal; [3] Unité Mixte de Recherche (UMR) 7257, Centre National de la Recherche Scientifique (CNRS) – Aix-Marseille Université, Architecture et Fonction des Macromolécules Biologiques (AFMB), Campus de Luminy, 163 Avenue de Luminy, 13288 Marseille CEDEX 09, France; [4] Institute for Molecular Bioscience, The Universy of Queenland, St Lucia, 4072, Australia.

* Equal contribution

Adapted from a manuscript accepted in Microbial Cell Factories.

Abstract

Animal venoms are large, complex libraries of bioactive, disulfide-rich peptides. These peptides, and their novel biological activities, are of increasing pharmacological and therapeutic importance. However, recombinant expression of venom peptides in *Escherichia coli* remains difficult due to the significant number of cysteine residues requiring effective post-translational processing. There is also an urgent need to develop high-throughput recombinant protocols applicable to the production of reticulated peptides to enable efficient screening of their drug potential. Here, a comprehensive study was developed to investigate how gene design, choice of fusion tag, location of expression, tag removal conditions and protease recognition site affect levels of expression and solubility of venom peptides produced in *E. coli*. The data revealed that expression of venom peptides in *E. coli* imposes significant pressure on cysteine codon usage suggesting that both *E. coli* codons need to be present at equivalent levels in genes designed *de novo* to ensure high levels of expression. Disulfide bond isomerase C (DsbC) was the best fusion tag for venom peptide expression, in particular when the fusion was directed to the bacterial periplasm. While the redox activity of DsbC was not essential to maximize expression of recombinant fusions proteins, redox activity did lead to higher levels of correctly folded target peptides. With the exception of proline, the canonical Tobacco Etch Virus (TEV) protease recognition site tolerated all other residues at its C-

terminus, confirming that no non-native residues, which might affect activity, need to be incorporated at the N-terminus of recombinant peptides for efficient tag removal. This study reveals that *E. coli* is a convenient heterologous host for the expression of soluble and functional venom peptides. Using the optimal construct design, a large and diverse range of animal venom peptides were produced in the micro-molar (µM) scale. These results open up new possibilities for the high-throughput production of recombinant disulfide-rich peptides in *E. coli*, an approach which is explored further in Chapter 6.

5.1. Introduction

Animal venoms comprise an arsenal of dozens to hundreds of structurally diverse disulfide-rich peptides that possess important pharmacological, therapeutic and biotechnological values. Considering the number of animal species that produce venoms and the average number of peptides per venom, the library of naturally evolved venom peptides may encompass millions of different molecules. These highly stable disulfide-reticulated peptides display formidable affinity and selectivity while presenting low immunogenicity making them attractive candidates for the development of novel therapeutics (Escoubas & King, 2009). In general, venom peptides target a variety of cell surface receptors, such as ion channels, and interaction with their molecular ligands dramatically affects cellular function (Lewis & Garcia, 2003). Venom peptides generally contain between 20 to 120 residues and include up to eight disulfide bonds that are critical for both biological activity and stability. Thus, the correct oxidation of cysteine residues leading to proper disulfide pairing is required for folding and functional activity. Unfortunately, the use of venom peptides as therapeutic or biotechnological molecules is still hampered by the difficulty to produce native and active proteins in sufficient amounts (Fernandes-Pedrosa, Félix-Silva, & Menezes, 2013).

De novo gene synthesis is the most convenient route to obtain genes for recombinant expression. This is particularly true for genes encoding venom peptides, as the sequence information from genomic and transcriptomic projects is usually not available as palpable DNA. Designing a gene to express a protein requires selecting from an enormous number of possible DNA sequences (Welch *et al*., 2011). In addition, gene family may affect effective gene design. For example, a high percentage of cysteine residues in venom peptides may impose particular constraints in levels of gene expression and these remain to be uncovered for the particular case of genes encoding animal venom peptides. Usually, gene design involves selecting a codon usage that maximizes levels of expression based on the codon bias of a subset of highly-expressed native host genes (Fuglsang, 2003; Henaut & Danchin, 1996). Expression may also be impaired by a strong mRNA secondary structure near the translational start site, inadequate GC content or presence of unwanted regulatory sequences recognized by the cellular expression machinery (Welch, Villalobos, *et al*., 2009). Although different studies have

analysed how genes can be designed efficiently, there is still no information about the major factors affecting expression of genes encoding reticulated peptides in heterologous hosts.

Escherichia coli is a highly robust bioreactor for heterologous protein expression. Several high-throughput platforms have been developed using this bacterium (Saez *et al*., 2014; Saez & Vincentelli, 2014). *E. coli* is particularly adequate to generate large libraries of recombinant proteins to apply to functional screens with biomedical and biotechnological relevance. However, production of disulfide-bonded proteins in bacteria is hampered by the lack of an effective post-translational system. Thus, in *E. coli* reticulated peptides are especially prone to aggregation or degradation due to possible mispairing of cysteine residues or undesirable intermolecular disulfide bonds. In addition, gene expression in bacteria is regulated by strong promoters, leading to the accumulation of recombinant proteins as insoluble aggregates or inclusion bodies. Different technologies have been developed to promote the correct oxidation of cysteine residues in recombinant proteins expressed in bacteria (Clement *et al*., 2015). Exporting the proteins to the *E. coli* oxidative periplasm is a well-established strategy although levels of recombinant protein can be limited by protein export (Nozach *et al*., 2013). For successful expression, two challenges must be met; (i) the peptide of interest must be maintained in a soluble state, and (ii) the correct disulfide bonds must be formed within the peptide. Recently some fusion tags displaying not only a solubilizing effect but also redox properties, such as the disulfide oxidoreductase (DsbA) and DsbC, were described by our group to enhance the solubility of venom peptides while promoting correct disulfide bond formation (Nozach *et al*., 2013; Saez *et al*., 2014). However, the most effective high-throughput compatible strategy to express a wide panel of correctly folded venom peptides in *E. coli* remains to be established.

Fusion tags are indispensable tools for protein expression and purification in bacteria (Costa, Almeida, Castro, & Domingues, 2014). However, presence of a fusion tag may interfere with protein function and its removal from the target protein is usually desirable. Tobacco etch virus (TEV) protease (Parks, Leuther, Howard, Johnston, & Dougherty, 1994) is one of the most popular enzymes used to remove fusion tags from recombinant proteins due to the stringent sequence specificity it displays. However, TEV protease may require a glycine (Gly) or serine (Ser) residue at the C-terminus (P1' position) of its recognition site (Dougherty, Carrington, Cary, & Parks, 1988), leaving a non-native Ser or Gly residues at the N-terminus of the target protein after tag removal. In the specific case of venom peptides it is well known that the N-terminal part of the peptide can contribute to the pharmacophore involved in receptor binding and thus the presence of an N-terminal fusion tag may affect biological activity (Karbat *et al*., 2007). Thus, removal of N-terminal tags is absolutely required to guarantee functional recombinant venom peptides. The FP7 European VENOMICS project is a consortium researching animal venoms to discover and develop innovative drugs. VENOMICS aims to establish a new paradigm in venom science by combining transcriptomics and proteomics to

explore the peptide content of venoms with high-throughput synthesis and recombinant expression of venom peptides to build large libraries of bioactive molecules for drug discovery. Within the VENOMICS project, we have examined and optimized gene design, the choice of fusion tag, as well as TEV cleavage conditions and recognition site, to improve the production of oxidized recombinant venom peptides in *E. coli*. Overall data reported here suggest that *E. coli* is an effective host to express milligram per litre quantities of correctly oxidized recombinant venom peptides using high-throughput technologies.

5.2. Materials and Methods

5.2.1. Design of gene variants encoding venom peptides

For the initial studies, 24 representative venom peptides originating from 21 different animal species were selected. The peptides had sizes ranging from 21 to 84 residues and contained between 2 to 7 disulfide bridges (see Table S5. 1 in Annex). The primary sequence of the 24 venom peptides was back-translated using ATGenium codon optimization algorithm, which uses a Monte Carlo repeated random sampling algorithm to generate three gene variants *per* peptide. This algorithm selects a codon for each position at a probability defined in a codon frequency lookup table. The lookup table applied to create the three variant designs of each gene varied in global codon usage within codons used preferentially in highly expressed or average native *E. coli* genes. Other factors considered for gene design were GC content, mRNA structure, absence of prokaryotic regulatory sequences and contiguous strings of more than 5 identical nucleotides, which were set not to vary within the different gene variants. Due to the use of Monte Carlo sampling for gene design, all the three variants were significantly different in sequence identity from each other. The average pairwise DNA sequence identity was 79.8% for the 24 datasets. Thus, in the initial phase of this work 72 genes were designed (3 gene variants of 24 peptides), which sequences are presented in Table S5. 1.

5.2.2. Gene synthesis, cloning and protein expression/purification of initial 72 variants

The 72 synthetic gene variants were produced using optimized procedures described in Chapter 3. The sequence coding for a TEV protease cleavage site (ENLYFQ/G) was engineered upstream of each gene. This sequence was identical for all 72 gene variants. Nucleic acids were synthesised containing Gateway recombination sites on each extremity. After PCR assembly, synthetic genes were directly cloned into pDONR201 using Gateway™ BP cloning technology (Invitrogen, USA) (Hartley, Temple, & Brasch, 2000b). Like for all the other plasmids and constructs used in this study, each construct was completely sequenced in both directions to ensure 100% consistency with the designed sequences. The 72 sequence entry clones were recombined using the Gateway™ LR cloning technology (Invitrogen, USA) to transfer the peptide-coding genes into pETG82A destination vector (Vincentelli *et al*., 2011).

Destination vector pETG82A contains the sequence encoding for a DsbC fusion partner, which is located at the 5´end of the inserted gene. All recombinant peptides fused with an N-terminal DsbC fusion tag contain an additional internal hexa-histidine (6HIS) tag for protein purification. Each variant plasmid was transformed into *E. coli* expression host strain BL21(DE3) pLysS (Invitrogen, USA). The choice of the plasmid and strain used in this experiment was based on our previous studies done on reticulated peptides (Nozach *et al*., 2013). Transformed cells were grown on solid media and resulting colonies were used to inoculate 4 mL of ZYP-5052 auto-induction medium (Studier, 2005a) supplemented with 200 µg/mL of ampicillin. Four independent colony isolates of each of the 72 recombinant strains were picked and cultured, so in total 288 cultures (72 recombinant strains grown in quadruplicate) were produced. All steps were carried out in 24 deep-well (DW24) plates following exactly the laboratory standard protocol (Klint *et al*., 2013; Nozach *et al*., 2013), which is described briefly below. ZYP-5052 medium is an auto-inducing buffered complex medium. Recombinant protein expression was induced following a standardized two-step process. Cells were grown at 37°C at 400 rpm in an orbital incubator shaker to quickly reach the glucose depletion phase just before the induction. After that step (4 hours, OD_{600nm} ~1.5), the temperature was lowered to 17°C for 18h to favour protein folding and soluble protein expression. Cells were collected by centrifugation at 2,500 $\times g$ for 10 min, re-suspended in 1 mL of lysis buffer (50 mM Tris, 300 mM NaCl, 10 mM Imidazole, pH 8.0, 0.25 mg/mL lysozyme) and the His_6-tagged recombinant proteins purified from crude lysates using an automated immobilized metal affinity chromatography (IMAC) procedure (Saez *et al*., 2014; Saez & Vincentelli, 2014). Briefly, the crude cell lysates were incubated with 200 µL of Ni^{2+} Sepharose chelating beads to capture the recombinantly expressed proteins, and then transferred into 96-well filter plates (Macherey-Nagel). The wells were washed twice with 1 mL of 50 mM Tris, 300 mM NaCl, 50 mM Imidazole, pH 8.0 buffer. The recombinant fusion proteins were eluted from the resin beads with 500 µL of elution buffer (50 mM Tris, 300 mM NaCl, 250 mM Imidazole, pH 8.0) into 96-deep-well (DW96) plates. All protein purification steps were automated on a Tecan liquid handling robot (Switzerland) containing a vacuum manifold. Analysis of the purified protein yields was performed on a Labchip GXII (Perkin Elmer, USA) microfluidic high-throughput electrophoresis system. This analysis provided an estimation of the molecular weight, purity and concentration of the proteins. All the quantitative values given in this manuscript are based on the calculation made by the Labchip GXII software.

5.2.3. Statistical analysis

Data related to yields of 24 purified recombinant fusion proteins originated from three different gene designs were subjected to ANOVA according to the general linear models procedure of SAS (SAS, 2004). The Least Squared Means procedure was used to detect significant

differences between high, medium and low expresser variants. Differences were considered significant when $P < 0.05$.

5.2.4. Construction of pHTP-derivative vectors to express venom peptides in *E. coli*

A collection of 5 novel vectors was constructed based on the prokaryotic expression vector pHTP1 (NZYTech, Portugal). The DNA sequences encoding a fusion protein tag were inserted into pHTP1 plasmid downstream of the T7 promoter, such that the protein tags would become fused to the N-terminus of the target peptide. DNA sequences encoding fusion tags were obtained by gene synthesis (see above) and included upstream and downstream NcoI restriction sites. Once inserted into pHTP1 backbone after digestion with NcoI, the five pHTP vectors conserved the C-terminal 6HIS tags for protein purification (see Table S5.2 in Annex). The five novel tags were based on disulfide-bond isomerase C (DsbC) and maltose-binding protein (MBP) sequences, some of the best tags for producing functional venom peptides in *E. coli* described to date (Anangi, Rash, Mobli, & King, 2012; Bende *et al*., 2014, 2015; Cardoso *et al*., 2015; Klint *et al*., 2013; Meng *et al*., 2011; Nozach *et al*., 2013; Saez *et al*., 2011; S. Yang *et al*., 2013). Thus, vector pHTP2 (pHTP-LLDsbC) encodes the sequence of DsbC for cytoplasmic expression, since it does not carry a signal peptide sequence (leader less-LL). In addition, pHTP3 (pHTP-mutDsbC) expresses a redox inactive mutant of DsbC, which includes two different mutations at the catalytic site (Cys100Ala and Cys103Ala), while in pHTP4 (pHTP-DsbC), the sequence of a signal peptide is included before the DsbC to allow export of the recombinant fusion protein to the periplasm. Similar vectors were also produced encoding MBP derivatives and were termed pHTP5 (pHTP-LLMBP) and pHTP6 (pHTP-MBP), respectively. The protein sequences of the six fusions created for this project are presented in Table S5.2. Schematic representations of the fusion proteins expressed from each vector are shown in Figure 5.3.

5.2.5. Cloning genes encoding 16 venom peptides into 6 pHTP vectors

The genes encoding 16 representative animal venom peptides were synthesised as described previously with a codon usage optimized for expression in *E. coli*. The 16 synthetic genes encoding venom peptides were directly cloned into pUC57. Upstream and downstream of all 16 genes, a 16 bp sequence was engineered to allow cloning into vectors of the pHTP-series using the NZYEasy cloning protocol (NZYTech, Portugal), which is based on a ligation independent cloning (LIC) method. Sequence and properties of the 16 genes produced here are presented in Table S5.3. The 16 different peptide genes were transferred from the pUC57 vector into each one of the 6 expression vectors in an experiment consisting of 96 cloning reactions. Reactions consisted of 240 ng of each linearized vector, 120 ng of the pUC57 derivative containing the target peptide gene, 1 µL of enzyme mix and 2 µL of 10x reaction

buffer. Cloning reactions were carried out in 20 µL final volume on a thermal cycler programmed as follows: 37ºC for 1 hour; 80ºC for 10 minutes and 30ºC for 10 minutes. The reaction mixtures were used to transform DH5α *E. coli* competent cells. Two colonies were picked for each construct and the presence of insert confirmed by PCR using the vector specific T7 and pET24a forward and reverse primers, respectively. All 96 plasmids containing the venom peptide genes were sequenced to confirm integrity of the cloned nucleic acid.

5.2.6. Recombinant expression and purification of TEV protease

The TEV derivative used in these studies is a mutant of TEV protease (TEV_{SH}), a kind gift of Dr Helena Berglund. This TEV variant is an improved version of TEV protease and was obtained after direct evolution studies. Thus, it presents high solubility, increasing levels of recombinant protease expressed in *E. coli* and an His tag to allow direct purification by IMAC (Van Den Berg, Löfdahl, Härd, & Berglund, 2006). The expression and purification of TEV_{SH} was done mainly following the published protocol (Van Den Berg *et al*., 2006) except for the LB medium that was replaced by ZYP-5052 (or Terrific Broth medium, TB) medium to reach a yield of purified TEV_{SH} up to 100 mg/L culture. At the end of the purification, the TEV_{SH} was dialyzed into 20 mM Hepes, 300 mM NaCl, 10% Glycerol (v/v), pH 7.4 to remove traces of DTT (DTT is part of buffer used to elute TEV protease), concentrated to 2 mg/mL and stored at -80ºC.

5.2.7. Recombinant protein expression and purification, and TEV cleavage protocol

The 96 recombinant pHTP derivatives (16 venom peptides × 6 pHTP vectors) were used to transform BL21(DE3) pLysS *E. coli* cells. Recombinant strains were grown in 4 mL of ZYP-5052 auto-induction medium supplemented with kanamycin (50 µg/mL). Recombinant strains were grown in DW24 plates at 37ºC for four hours in a microplate shaker. The temperature was then dropped to 17ºC and cells were left to grow for 16-20 hours. Cells were harvested by centrifugation of the DW24 plates at 2,500 ×*g* for 10 min. Recombinant peptides fused with different tags were purified as described above (see 5.2.2 section). After purification through IMAC, recombinant toxins were digested with TEV_{SH} protease. The TEV cleavage protocol used here to remove fusion tags from recombinant peptides was as described elsewhere (Saez & Vincentelli, 2014). The cleavage was performed overnight at 30ºC (or 4ºC when stated) with a protein/TEV ratio of 1:5 or 1:10 (w/w). When necessary, the cleavage buffer was supplemented with fresh DTT. The cleavage efficiency was calculated as previously described (Kapust, Tözsér, Copeland, & Waugh, 2002).

5.2.8. Tag removal and liquid chromatography-mass spectrometry (LC-MS)

Detection of oxidized recombinant peptides was performed using liquid chromatography-mass spectrometry (LC-MS). After overnight TEV cleavage of purified fusion peptides, samples were acidified for 1h with a solution of 5% acetonitrile (ACN) and 0.1% formic acid. Precipitated material (TEV protease, fusion tags and misfolded peptides) was removed by centrifugation at 4,100 x*g* for 10 minutes. Aliquots (20 µL) of the 96 cleaved samples were analysed on a C18 reverse phase column at 37°C (Hypersil GOLD column, 50 x 1.0 mm, 1.9 µm, 175 Å, ThermoScientific), with a flow rate of 200 µL/min on an UHPLC-MS with electrospray ionization (Accela High Speed LC system with detector MSQ+, ThermoScientific, San Jose, CA). The gradient slope (solvent A: water, B: acetonitrile, both solvents containing 0.1 % formic acid) went from 5 to 40% B in 2 min followed by an 80% wash and re-equilibration (total time: 6 minutes). Mass Spectrometry acquisition was performed in the positive ion mode from *m/z* 100 to 2000. To confirm correct peptide molecular weight, the resulting mass spectra were de-convoluted using manual calculations. The isotopic pattern measured was compared with the theoretical one determined from the amino acid sequences using Data Explorer software (Version 4.9, Applied Biosystems). The quantitative calculation of peptide yields was determined using automatic processing with Xcalibur software (ThermoScientific), by OD_{280nm} measurement and peak areas integration.

5.2.9. Generation of N-terminal variants of DNA/RNA-binding protein Kin17

To test the efficacy of TEV protease to cleave peptide chains including variations at the C-terminus of the consensus recognition site of the enzyme (ENLYFQ/X), the gene encoding the C-terminal domain of the DNA/RNA-binding protein Kin17 (Kin17) from *Homo sapiens* was synthetized. PCR was used to create 20 gene variants encoding derivatives of the Kin17 protein with 20 different N-terminal amino acids at the TEV recognition site. The genes were produced by PCR including the reverse primer HSr, 5´-GGGGACCACTTTGTACAAGAAAGCTGGGTCTTATTAAAGTTTAGAGATGTCTTCAT-3´ and the forward primers presented in Table S5.4. Amplified nucleic acids contained Gateway recombination sites on each extremity. Thus, genes were initially directly cloned into pDONR201 using Gateway™ BP cloning technology (Invitrogen, USA). The 20 gene variants were subsequently cloned into pDest17 vector (Invitrogen, USA) using Gateway™ LR cloning technology (Invitrogen, USA). Resulting expression plasmids encoded for Kin17 derivatives containing an N-terminal 6HIS tag and a TEV recognition site combining 20 variations at the residue occupying its C-terminal (P1') position. Primary sequence of both proteins and respective genes are presented in Table S5.4. The 20 plasmid derivatives were used to transform BL21(DE3) pLysS *E. coli* cells and recombinant proteins were produced, purified and cleaved as described above.

5.3. Results

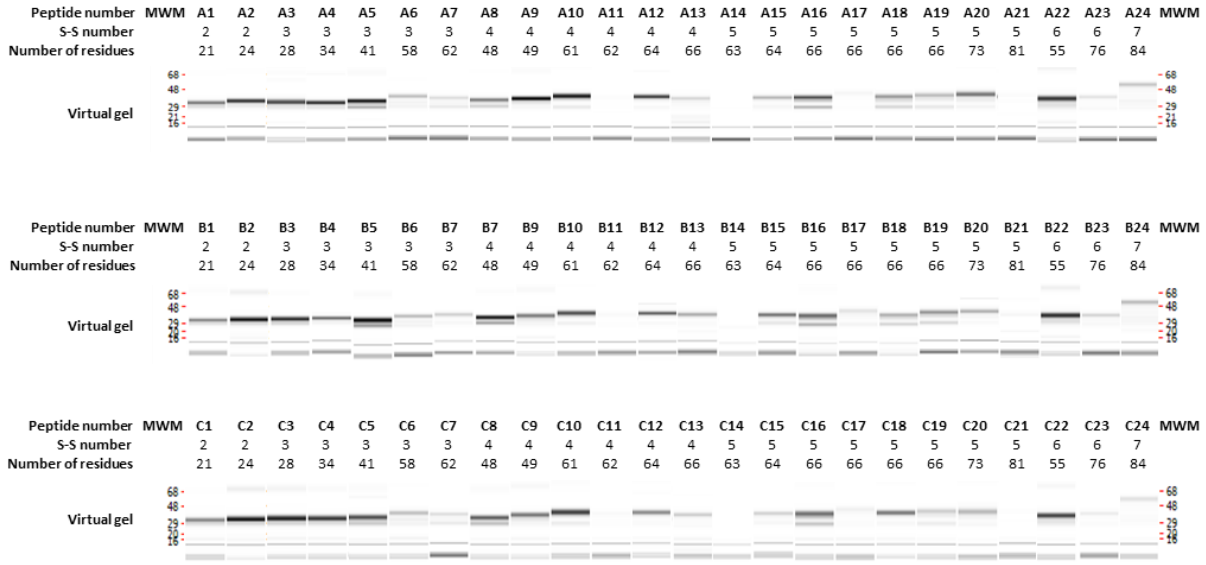
5.3.1. Codon usage of venom peptide encoding genes cause expression differences

Twenty-four genes of various lengths encoding venom peptides from different species and containing different numbers of disulfide bridges were chosen to explore the effects of codon usage on soluble levels of purified proteins. These genes encode venom peptides that are evolutionarily, structurally and functionally diverse. The experiment aimed to evaluate if subtle changes in codon usage can affect levels of recombinant peptide expression in *E. coli*. Three variants of each gene (variants A, B and C) were initially designed by back-translating venom peptide sequences using a Monte Carlo repeated random sampling algorithm to select codons probabilistically from codon frequency lookup tables. The codon usage of the 72 devised genes (3 variants of 24 genes) is presented in Table S5.5 and reflects the codon usage of *Escherichia coli* genes at moderate to high levels. However, created gene variants incorporated changes in primary sequences that reflect the random sampling of codon selection and the overall freedom permitted by the algorithm used for gene design. Thus, the average pairwise DNA sequence identity was 79.8% within the three variants of the 24 datasets. The 72 genes were synthesised and cloned using the Gateway™ system into pETG82A prokaryotic expression vector under the control of a T7 promoter and in fusion with the gene encoding the DsbC fusion tag for cytoplasmic expression. *E. coli* BL21(DE3) pLysS were transformed with the 72 plasmids and grown in quadruplicate using auto-induction media. Fusion proteins were purified and protein integrity and yield measured by Caliper analysis (Figure 5.1, A). Depending on the peptide of interest, the purified fractions run on the Caliper mainly as a single band (peptides 1, 2, 3, 4…) or as a double band (5, 8, 16, 18…). The single band represents the good protein population (DsbC-His-peptide) while the lower band (around 29 kDa) corresponds to the DsbC-His protein alone after truncation/degradation of the target peptide. This lower band probably indicates that there is a portion of the peptide population that was not properly folded and was degraded during the expression or the purification processes. The protein concentration depicted in Figure 5.1, B has been calculated by integrating only the DsbC-His-peptide band using the Caliper LabChip software. The data, presented in Figure 5.1, B revealed that yields of purified fusion protein varied from ~1 mg/L (for fusion protein 11) to above 100 mg/L (for fusion proteins 4, 5, 8, 9, 10 and 18). For the vast majority (19/24), the quantities of fusion protein purified allowed the purification of milligram scale of target peptide per litre of culture (assuming a cleavage and purification yield around 100%) while for the remaining five peptides (11, 14, 17, 21 and 23) a larger volume of culture would be needed. The correlation between primary sequence of gene variants and properties that have been suggested to affect expression was analysed. Deleterious motifs, such as 5′ mRNA secondary structures could not have affected levels of expression as venom peptide genes were all fused to the same 5´-prime sequence, which encodes the protein fusion tag. There was no correlation between

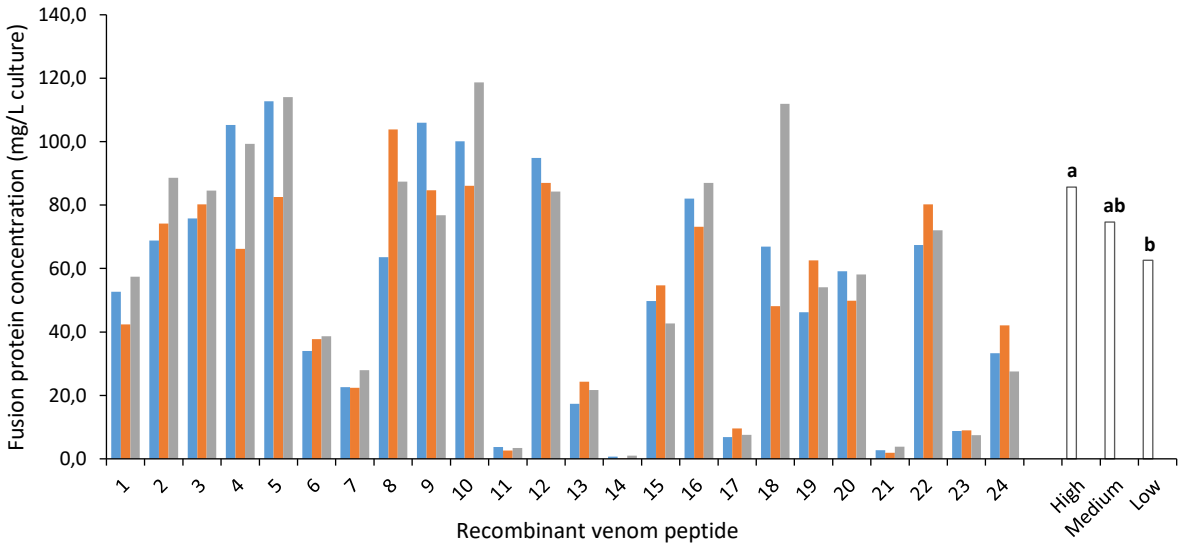
protein expression and number of disulfide bridges, peptide size, CAI value and GC content (data not shown). This suggests that differences in gene expression were determined by other sequence related properties in particular by codon usage.

Figure 5.1| Yields of 24 purified recombinant fusion proteins originated from 3 different gene designs.

A.



B.



The proteins were expressed as DsbC-His fusions in the cytoplasmic compartment of *E. coli* cells. Panel A: Virtual gel showing the expression levels of 24 recombinant peptides obtained from the gene design A, B and C that were purified through IMAC and evaluated using the Labchip GXII (Caliper, USA). Panel B: Comparison of expression levels of variant A (blue), variant B (orange) and variant C (gray) of the 24 recombinant peptides. On the right are represented the means of high, medium and low expresser variants calculated for the 16 fusion peptides produced at higher yields. Means without a common letter differ at $P<0.05$.

In order to investigate how changes in codon usage affected levels of recombinant peptides, the relation between protein yields of the low, medium and high expresser variants within the

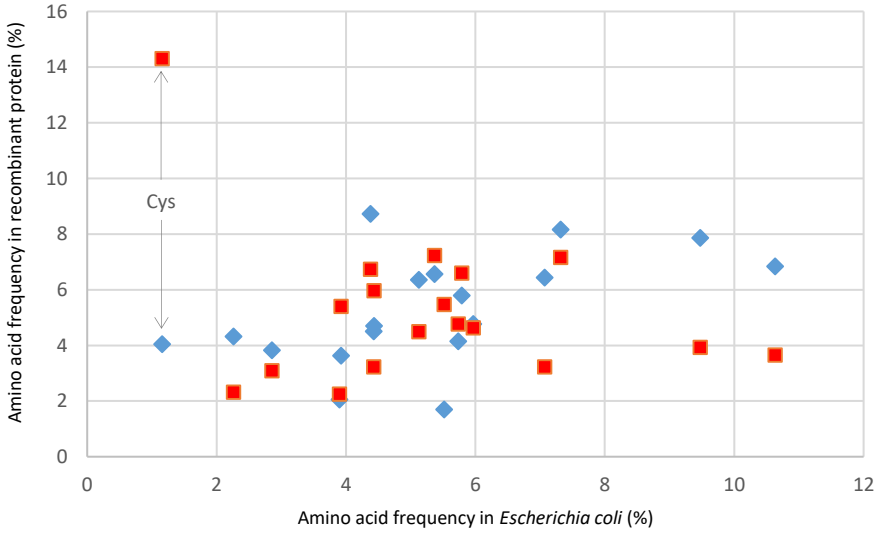
24 data sets were compared. Fusion proteins expressing at lower levels, peptides 7, 11, 13, 14, 17, 21, 23 and 24 (Figure 5.1), were excluded from the analysis. The data revealed that protein yields of the high, medium and lower expresser variants of the peptides analysed were significantly different. Thus, lower expressers produced on average 62.6 mg/L of recombinant fusion protein, while fusion protein yields of higher expressers were, on average, of 85.7 mg/L (Figure 5.1, B). These differences are significantly different (*p-value* = 0.013). To evaluate what differences in codon usage could explain observed differences in protein expression, the codon usage of low and high expressing variants were compared. Codon usage tables including genes containing the fusion tag are presented in Table S5.6. Major differences in codon usage concern in particular one amino acid, cysteine, although slight changes were also observed for other residues, in particular arginine, asparagine, glutamate, histidine, isoleucine, phenylalanine and serine. Summary codon usage data for these 8 amino acids is shown in Table 5.1.

Table 5.1| Codon usage of genes encoding high and low expresser variants (HE and LE, respectively) encoding either venom peptides or their respective fusion proteins. Codon frequency is shown as Fc.

Amino acid	Codon	*Fc*, HE, peptides	*Fc*, LE, peptides	*Fc*, HE, fusion	*Fc*, LE, fusion	*Fc, E. coli*
Arginine (Arg)	AGA	0	0	0	0	0.07
	AGG	0	0	0.17	0.17	0.04
	CGA	0	0	0	0	0.07
	CGC	0.38	0.44	0.35	0.39	0.36
	CGG	0	0	0	0	0.11
	CGT	0.63	0.56	0.48	0.45	0.36
Asparagine (Asn)	AAC	0.55	0.64	0.47	0.49	0.51
	AAT	0.45	0.36	0.53	0.51	0.49
Cysteine (Cys)	TGC	0.49	0.58	0.49	0.56	0.54
	TGT	0.51	0.42	0.51	0.44	0.46
Glutamate (Glu)	GAA	0.84	0.76	0.62	0.60	0.68
	GAG	0.16	0.24	0.38	0.40	0.32
Histidine (His)	CAC	0.26	0.48	0.33	0.35	0.43
	CAT	0.74	0.52	0.67	0.65	0.57
Isoleucine (Ile)	ATA	0	0	0	0	0.11
	ATC	0.49	0.32	0.57	0.54	0.39
	ATT	0.51	0.68	0.43	0.46	0.49
Phenilalanine (Phe)	TTC	0.28	0.44	0.14	0.16	0.42
	TTT	0.72	0.56	0.86	0.84	0.58
Serine (Ser)	AGC	0.51	0.40	0.67	0.65	0.25
	AGT	0.09	0.16	0.11	0.13	0.16
	TCA	0.05	0.05	0.11	0.11	0.14
	TCC	0.11	0.16	0.07	0.08	0.17
	TCG	0.13	0.15	0.02	0.02	0.14
	TCT	0.11	0.07	0.02	0.01	0.15

The codon bias observed for low expressers genes revealed a preference for Cys-TGC codon while in high expressing genes Cys-TGT is favoured. In addition, in low expressing genes Cys-TGC is used 1.38 times more frequently than Cys-TGT, while in high expressing genes Cys-TGT is only used 1.04 more often than Cys-TGC. This observation suggests that high expression of genes encoding peptides requires a similar contribution of both Cys-TGC and Cys-TGT codons, suggesting that a higher percentage of one codon compared to the other will affect expression. Cysteine codon usage in *E. coli* also points to a balanced utilization of the two codons (Table 5.1). To investigate factors that may explain this observation, amino acid frequency in *E. coli* genes and within the 24 venom peptides selected for this study and their associated fusion proteins were compared. The data, presented in Figure 5.2, revealed that cysteine is ~12.5 and 3.5 times more frequent in venom peptides (14.3%) and in the recombinant fusion proteins (4.1%), respectively, than in *E. coli* (1.16%). Thus, the expression of venom peptides at high levels may be promoted by the presence of the two cysteine codons at similar frequency in the synthetic genes to avoid the depletion of one codon when genes are expressed at very high levels.

Figure 5.2| Comparison of amino acid frequency in *Escherichia coli* with the frequency of each amino acid in recombinant peptides analysed in this study.



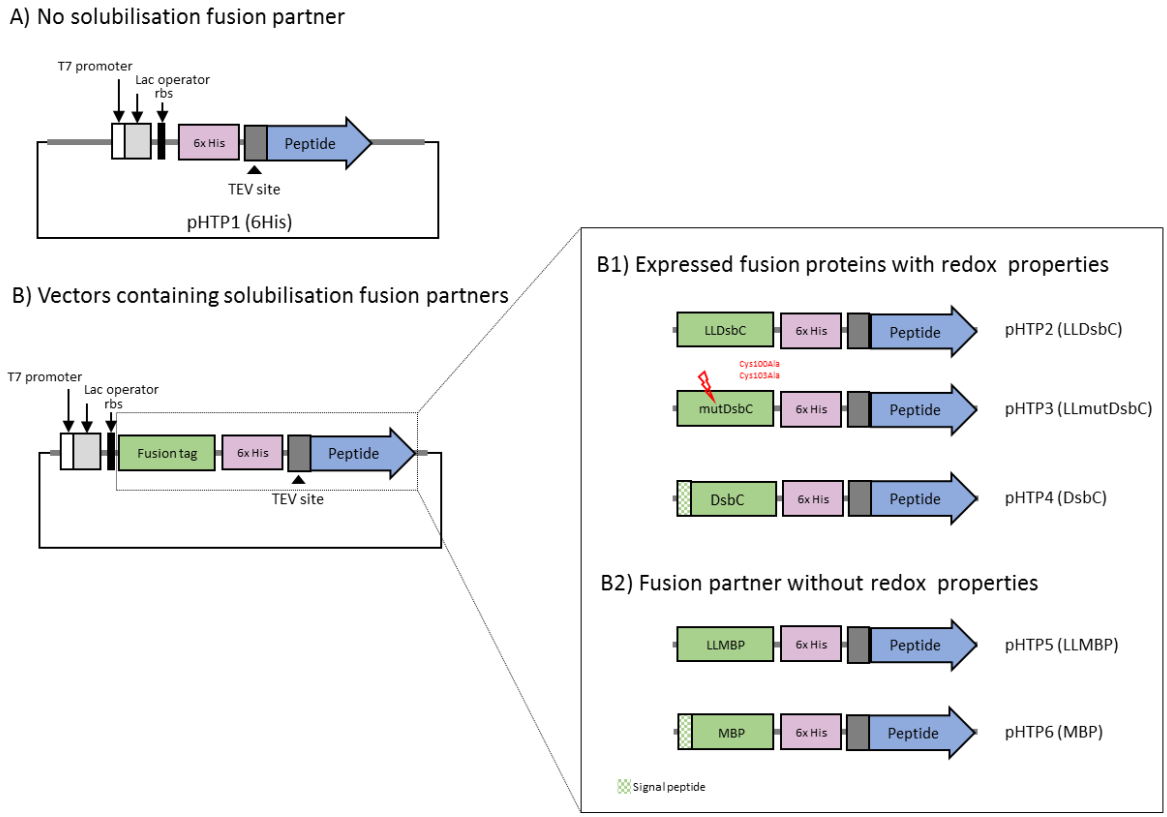
Percentage of abundance of each amino acid in fusion DsbC proteins is displayed in blue. In red, frequency of the same amino acid in venom peptide encoding genes analysed in this study excluding the sequence encoding the fusion tag.

5.3.2. Levels of expression of venom peptides are affected by the fusion tag

Five novel vectors for recombinant protein expression in *E. coli* were constructed by inserting different fusion tags into the pHTP1 backbone. All fusion tags are to be inserted at the N-terminus of the recombinant peptides (Figure 5.3). Two of the vectors encode fusion partners that contain a signal peptide (leader sequence) to target venom peptide expression into the periplasm (pHTP4, pHTP6). The remaining fusion tags will lead to cytoplasmic recombinant

protein expression (Figure 5.3). In all cases a 6HIS tag was introduced to enable the downstream purification of the fusion proteins using immobilized metal affinity chromatography (IMAC). A TEV protease cleavage site (ENLYFQ/G) was introduced in all synthetic genes to enable removal of the fusion partner. In addition to the 6HIS affinity tag alone (pHTP1), the 5 novel vectors include the disulfide isomerase DsbC or the maltose binding protein (MBP). An inactive mutant derivative of DsbC, which contains two amino acid changes at the catalytic site, was produced to try to discriminate the roles of DsbC in passive solubilisation (relating to fusion protein yield) and redox activity (relating to the yield of correctly folded target peptide). The schematic representation of all vectors use in this study is presented in Figure 5.3.

Figure 5.3| Schematic representation of the expression vectors that contain fusion tags with and without redox properties, which were used for cytoplasmic and periplasmic expression of venom peptides in *Escherichia coli*.

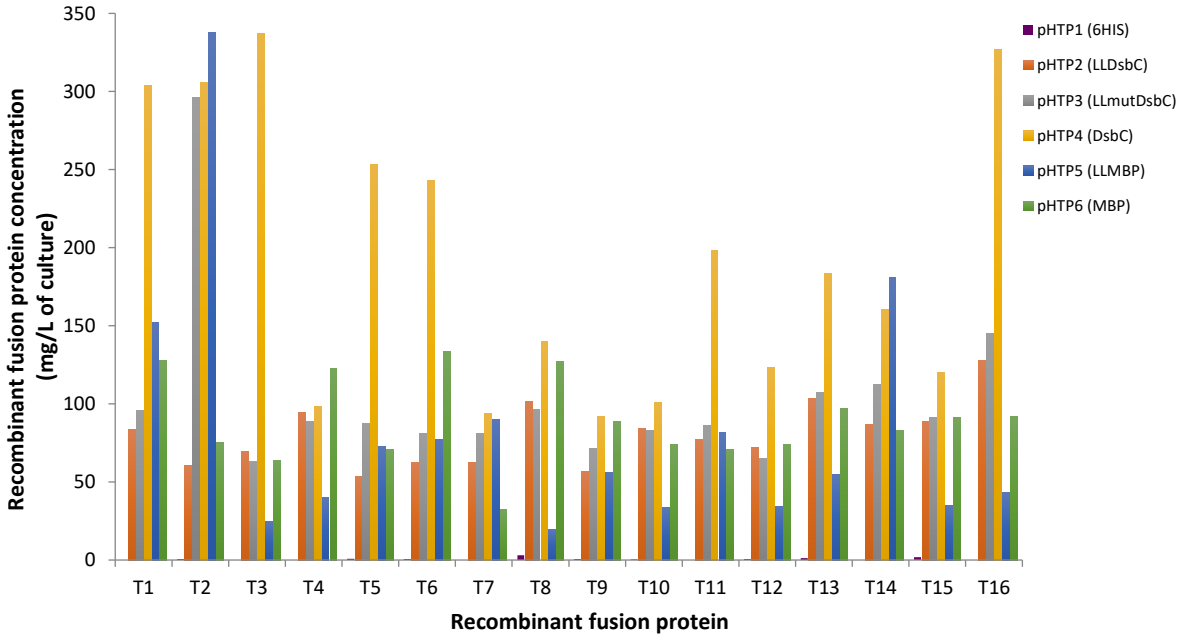


All vectors include a T7 promoter, a ribosome binding site (rbs), a lac operator, a 6HIS tag for nickel affinity purification and a Tobacco Etch Virus (TEV) protease cleavage site. The 6HIS tag is N-terminal for pHTP1 vector (A) and internal for expression vectors including fusion tags (B). pHTP4 (DsbC) and pHTP6 (MBP) carry fusion tags containing a signal peptide (represented in crossed green lines) to target exportation of the fusion protein to the periplasm of *E. coli* cells. The inactive DsbC fusion partner, which contains two mutations at the catalytic site (Cys100Ala and Cys103Ala), was inserted into pHTP3 (LLmutDsbC). LL, leader less (fusion tag with no signal peptide).

Sixteen well characterized venom peptides with different origins and representing different folds and cysteine bond patterns were selected for this study (see Table S5.3 in Annex). The

16 synthetic genes were inserted into the six different expression vectors (see Table S5.2 in Annex and Figure 5.3) generating a total of 96 recombinant plasmids. The 96 constructs were transformed in BL21(DE3) pLysS cells. Recombinant *E. coli* strains were grown in auto-induction media to obtain high cell densities. After nickel affinity purification, systematic analysis of the Labchip GXII electropherograms was performed to determine the concentration of the purified proteins and compare the apparent molecular weight of the purified fusion proteins with their expected theoretical molecular weight. Data, presented in Figure 5.4, revealed that the 16 peptides can indeed be produced using a fusion tag. Depending on the peptide and the fusion used, the levels of purified fusion proteins varied from zero (mostly when peptides were cloned in pHTP1) to more than 300 mg of purified fusion protein per litre of culture. Overall, smaller peptides seemed to be easier to produce than larger ones but there are several counter examples (like T16 which is the largest peptide of the study).

Figure 5.4| Yields of 96 purified recombinant fusion proteins originated from 16 different animal venom peptides in 6 fusions.



Peptides are organized by increasing mass. Each fusion is represented by a colour code. Yield is expressed in milligram of fusion per litre of culture. Fusion proteins were purified through IMAC and evaluated using the Labchip GXII (Caliper, USA).

For peptides T2, 4, 7, 8, 9, 14 and 15, different vectors seem to be appropriate for fusion protein expression but in most cases (13 out of 16) vector pHTP4 (DsbC) outperformed all other vectors. In contrast, without exception, soluble expression with 6HIS tag alone was always very low. The presence of the signal peptide lead to higher levels of expression for DsbC in all cases (pHTP4 versus pHTP2) doubling on average the amount of expressed fusion protein, while for the case of MBP (pHTP6 versus pHTP5) its presence gives a similar trend

(10/16). Thus, when MBP was used as a fusion tag, periplasmic folding did not favour yield and there are even two cases where the cytoplasmic MBP lead to a better outcome (for peptides T2 and T14). Finally, the inactivation of DsbC biological function had no effect in expression levels of venom peptide fusion proteins, as expression of pHTP2 and pHTP3 were, in general, very similar.

5.3.3. Fusion cleavage, peptide yield and correct oxidation state is mainly affected by the fusion partner and DTT concentration in TEV cleavage buffer

Optimal TEV cleavage conditions to release target peptides from fusion tags are affected by several parameters including enzyme/substrate ratio, buffer composition, incubation period and temperature. To investigate which conditions would lead to the best yield of folded venom peptides, eight peptides were selected from the list of 16 peptides produced in the previous experiment (see Table S5.3, peptides in italic). Because the DsbC fusion partner outperformed other tags in terms of fusion protein yields and general applicability, these constructs were selected for the TEV cleavage optimization study. In order to simplify the study, several parameters were kept constant for all the tests (based on previous in-house experiments (Nozach *et al*., 2013; Saez *et al*., 2014; Saez & Vincentelli, 2014)): the concentration of purified fusion protein (1 mg/mL), a fusion/TEV ratio of 1/10 (w/w), the buffer composition, the temperature (30°C) and the incubation period (18h). The cleavage buffer chosen was the IMAC protein elution buffer (50 mM Tris, 300 mM NaCl, 250 mM Imidazole, pH 8.0). This allows optimization of cleavage conditions using elution buffer, which considerably simplifies downstream processing. The only parameter that was fine-tuned in this experiment was DTT concentration present in the cleavage buffer, which varied from 0 to 2mM (0; 0.1; 0.5; 2 mM fresh DTT). Indeed, while the TEV protease requires reducing conditions for optimum cleavage, an excessive concentration of DTT could lead to the reduction of the peptides' intra-disulfide bridges and loss of folding and biological activity. After an 18h incubation period, an aliquot of the TEV-fusion peptide mixture was acidified. A fraction ("before" versus "after" TEV cleavage) was loaded on the Caliper to determine the cleavage efficiency and a sample was analysed on the liquid chromatography-mass spectrometry (LC-MS) to confirm the oxidation state. Cleavage efficiency, mass analysis and peptide yields are summarized in Figure 5.5. Cleavage occurred in the 32 conditions tested in the assay (ranging from 30% to 100% efficiency). As expected, cleavage was not complete in the absence of DTT and complete with the highest concentration of DTT (2 mM). Out of eight peptides, seven could be detected oxidized in mg quantities per litre of culture. From the 32 samples, 19 gave the correct oxidized mass on the LC-MS with various yields depending on the DTT concentration during cleavage. For the highest concentration of DTT (2 mM), all peptides were denatured. From the seven peptides correctly detected, four gave the highest recovery in 0.1 mM DTT, while two needed no DTT and finally one needed 0.5 mM DTT for optimum recovery. A DTT concentration of 0.1

mM appeared to be the best compromise to be kept for the following experiments and the production pipeline of VENOMICS project (Chapter 6). Aliquots of the 7 peptides was concentrated to 2 and 4 mg/mL and subjected to the same experiment with 0.1 mM DTT to confirm that cleavage would be possible in these conditions. On average, the efficiency dropped by 20% but occurred in all cases (data not shown).

Figure 5.5| TEV cleavage efficacy in various concentrations of DTT.

Protein	Number of disulfide bridges	Number of amino acids	Cleavage efficiency (%)				Oxidized peptide detected (mg/L of culture)			
			DTT concentration (mM)				DTT concentration (mM)			
			0	0.1	0.5	2	0	0.1	0.5	2
T1	3	34	90	90	90	100	13.5	17.6	16.0	-
T2	3	41	40	50	95	100	5.5	12.0	21.0	-
T3	3	55	70	100	100	100	8.6	4.0	2.4	-
T4	4	41	30	80	90	100	4.2	6.4	4.0	-
T6	4	65	30	50	70	100	1.5	3.9	-	-
T8	5	55	50	100	100	100	7.0	5.6	3.6	-
T15	7	83	80	100	100	100	1.5	3.7	-	-
T16	7	84	80	90	100	100	-	-	-	-

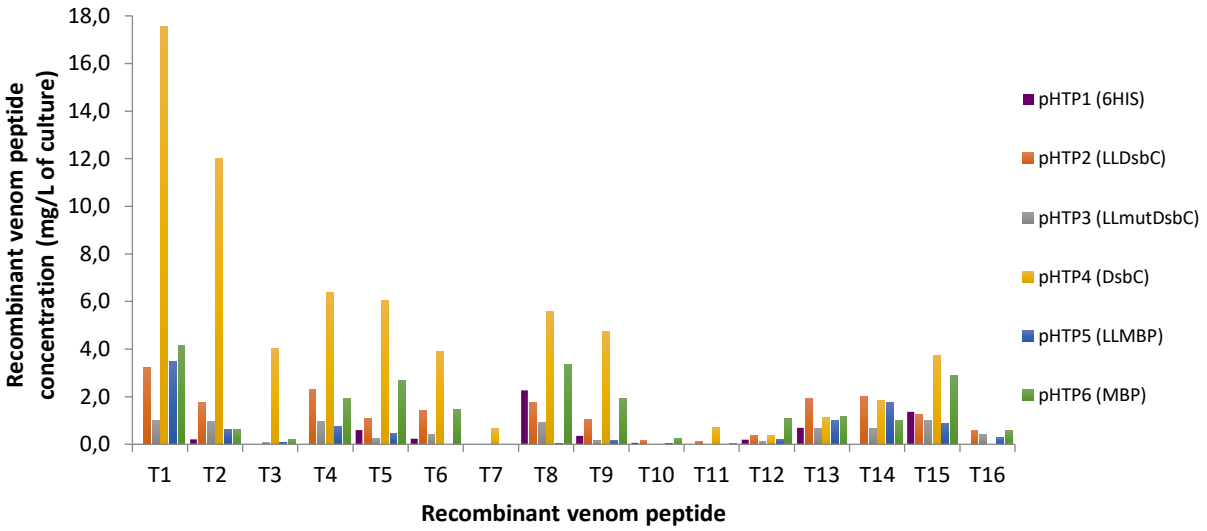
The cleavage efficiency represents the percentage of fusion protein cleaved for each DTT concentration (0 to 2 mM) quantified by Labchip GXII (Caliper, USA) and depicted in percentages. The correct oxidation state of the purified peptide was confirmed by LC-MS (green: mass corresponds to oxidized peptide, red: no peptide or no correct mass detected). When a correct mass is detected the yield of peptide per litre of culture quantified by integration of the LC-MS peaks is indicated in the well.

Following the optimization of cleavage conditions, the 96 purified fusion proteins (16 in 6 vectors, see Figure 5.4) were cleaved in the presence of 0.1 mM DTT at the concentration of the purified pools (ranging mostly from 0.2 mg/mL to 2 mg/mL) following the protocol described above. After cleavage and acidification, an aliquot of the 96 samples was analysed by LC-MS to confirm the correct molecular mass, yield and the oxidation state of the final recombinant venom peptide. When detected, the 96 recombinant peptides had molecular masses in agreement with the expected masses given fully oxidized cysteine residues; the reduced forms of the proteins were never detected (data not shown), probably due to precipitation of the incorrectly oxidized peptides during cleavage and acidification steps. The final yield for the 96 constructs, presented in Figure 5.6, was expressed as absolute peptide final yield in mg/L culture or normalized to 100% for each peptide relative to the vector used for expression. The data (Figure 5.6) revealed that all 16 peptides under study could be produced recombinantly but at different levels. Similarly to what was observed for yields obtained for the fusion variants (Figure 5.4), a general trend suggests that the shorter peptides are easier to produce than the

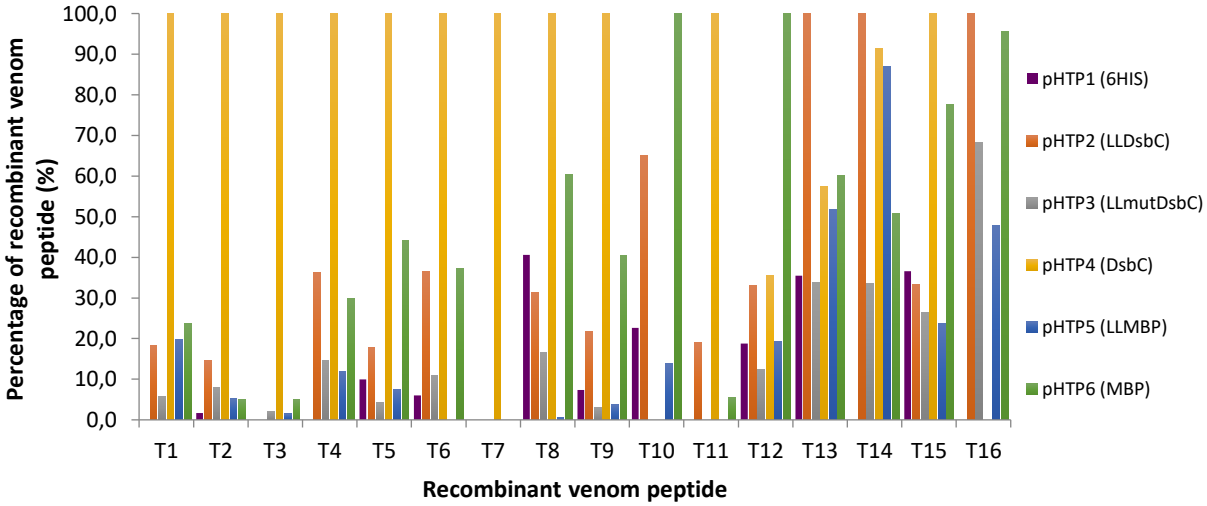
longer ones. The final venom peptide yield varied greatly with a fifty-fold difference between the worst case (0.3 mg/L, T10 in pHTP6) and the best case (17.6 mg/L, T1 in pHTP4). The dataset also revealed that there was a significant drop in yields following fusion tag cleavage. This is probably due to the harsh recovery condition after cleavage (acidification in 5% acetonitrile, 0.1% formic acid) where any misfolded peptide precipitates.

Figure 5.6| Yields of 96 purified recombinant peptides after tag removal.

A)



B)



Peptides are organized by increasing mass. Each original fusion tag used to express each peptide is represented by a colour code (identical to Figure 5.4). Yield is in milligram of oxidized peptide per litre of culture. The correct oxidation state of the purified peptide was confirmed by LC-MS and the yield of peptide per litre of culture quantified by integration of the LC-MS peaks. A) Concentration of recombinant peptide in mg/L of culture. B) Yield of peptide is presented in percentage relative to the best condition to visualize better the low expressing peptides.

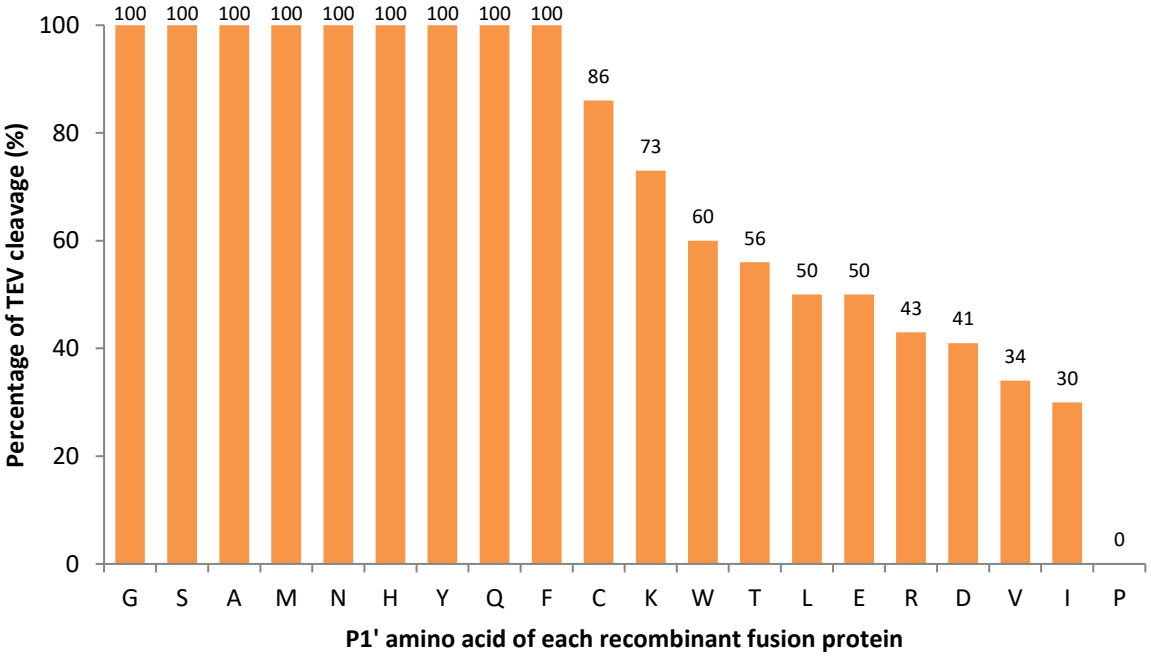
As expected from the fusion yields, even if recovery is high, peptides produced with pHTP1 gave the lowest quantities of peptides overall (0.4 mg/L on average, with a maximum of 1.9 mg/L for T8). In contrast, peptides produced with pHTP4 (DsbC) reached the best final peptide yields for 11 of 16 peptides, and of these (with the exception of T11) yields were more than 2 mg/L for each peptide (4,6 mg/L average). Overall, DsbC fusions (for either periplasmic or cytoplasmic expression) successfully produced 14 out of 16 venom peptides. Furthermore, one peptide (T7) could only be produced in the periplasm, using the DsbC fusion partner, from the pHTP4 vector. For peptides produced preferentially from other vectors (T10, 12, 13, 14, 16), yields do not surpass 2 mg/L, highlighting the robust expression from the pHTP4 vector. For these five peptides, highest yields were achieved with cytoplasmic expression and a DsbC fusion partner in three cases, followed by periplasmic expression with the MBP fusion partner in two cases. In all cases (except T14), the fusion exported to the periplasm (for either for DsbC or MBP) outperformed its cytoplasmic equivalent, indicating that at least part of the folding could occur in the periplasm and part could occur *ex vivo* during the purification. A striking demonstration that DsbC (and probably MBP) acts, in part, as a passive solubilising agent inside the bacteria is the fact that while the fusion yields between the DsbC and the mutated DsbC constructs were very similar for most peptides (Figure 5.4). After cleavage and recovery, the overall yield of active peptide is on average three times higher in the case of the redox-active DsbC fusion than with its mutated DsbC counterpart.

5.3.4. The nature of the C-terminal (P1') residue of the TEV cleavage site does not significantly affect cleavage efficacy

The N-terminus of some venom peptides can contribute to their receptor binding sites. Thus, it is possible that the introduction of a single extra residue at the N-terminus of the peptide may affect its biological activity (Karbat *et al*., 2007). The canonical TEV protease recognition site requires a Gly or Ser residue at its C-terminus (P1' position), leaving a non-native Ser or Gly residues at the N-terminus of the target peptide after tag removal. A previous study (Kapust *et al*., 2002) suggested that, with the exception of proline, all the amino acid side-chains could be accommodated in the P1´ position of the TEV protease recognition site with little impact on the efficiency of processing. The analysis was, however, performed in optimal TEV buffer conditions. In order to be time-efficient, the venom peptide production pipeline can not accommodate additional steps such as buffer exchange into optimal TEV conditions. Thus, the capacity of TEV protease to act in the IMAC elution buffer (50 mM Tris, 300 mM NaCl, 250 mM Imidazole, 0.1 mM DTT, pH 8.0), a non-optimum buffer for TEV proteolysis, was investigated. The TEV_{SH} protease used in this study (Van Den Berg *et al*., 2006) was selected because it is easy to overproduce and purify in *E. coli* at very high quantities (up to 100 mg/L culture). However, the cleavage specificity of this recombinant derivative of TEV protease remains unknown, in particular when various amino acids occupy the P1´ position of its

recognition site (Dr. Helena Berglund, personal communication). Thus, to explore the activity of this TEV_{SH} protease in non-optimal conditions and when the P1´ position of the protease recognition site is varied, 20 test-cleavage fusion protein sequences were produced. Each fusion protein contained an N-terminal 6HIS tag, an internal TEV recognition sequence, each containing a different amino acid at the P1´ position, fused C-terminally to a truncated form of the DNA/RNA-binding protein Kin17 from *Homo sapiens*. Proteins were purified and subjected to TEV protease cleavage in the same conditions as those used to cleave the 96 fusion tags (see above). The data, presented in Figure 5.7, confirm previous data collected by Kapust and colleagues (2002) and suggest that, with the exception of proline (the probability of having a proline in position 1 of naturally-sourced venom peptides is low), all other residues can be accommodated at the P1´ position of the TEV protease recognition site while retaining some cleavage activity. However, consideration should be given to peptides with an N-terminal Trp, Thr, Leu, Glu, Arg, Asp, Val or Ile, where cleavage efficiency dips to 60% or less. In these cases, a compromise between cleavage efficiency/production yield may need to be reached, depending on how well the peptide expresses. Note that in contrast to venom peptides, Kin17 does not contain any cysteine residue. Thus, the successful cleavage yield when a cysteine residue is in position 1 (86%) obtained with Kin17 needs to be confirmed for the cases of peptides with a cysteine in position 1 that would probably be involved in a disulfide bridge in the native protein.

Figure 5.7| TEV protease cleavage efficiency of Kin17 with 20 different amino acids located at position P1' of the recognition site.



The amino acids are organized from the easiest to the most difficult ones to cleave. Values are represented in percentages.

5.4. Discussion and conclusions

Today *de novo* gene synthesis is replacing the classic cloning approaches for the construction of transgenes and thus it is critical to develop effective gene design algorithms that could sustain high levels of heterologous gene expression (Czar, Anderson, Bader, & Peccoud, 2009). Effective design methods may require attending to particular properties of different protein families. Thus, the intrinsic abundance of cysteine residues in venom peptides is potentially a critical factor that may affect the recombinant expression of such peptides. Here, by designing and synthesising 72 individual genes encoding 24 different peptides, nucleic acid sequence differences affecting the levels of soluble expression of venom peptides were identified. High expresser gene variants produced up to twice the levels of recombinant protein when compared with low expressing ones. Factors affecting expression levels were identified by comparing the codon usage of high and low expresser variants. The data revealed that most of the variation in expression can be explained, primarily, by differences in the frequency of cysteine codons but also, at a lower level, for arginine and isoleucine. Thus, data presented here reveal that high levels of expression of venom peptides require a similar usage of the two cysteine codons Cys-TGT and Cys-TGC. It is now well established that high translation rates contribute to deplete the cellular translational machinery (Dong, Nilsson, & Kurland, 1995). Considering the levels of expression of the heterologous proteins reported in this study we estimate that between 25-40% of the total proteins produced by the bacterial cell comprise recombinant fusion polypeptides. Overexpression of recombinant genes in *E. coli* leads to a significant change in the amino acids being used for protein synthesis in favour of the recombinant protein. In the particular case of venom peptides, cysteine is a highly frequent residue being around four times more frequent in the recombinant fusion genes than in *E. coli* native proteins. Thus recombinant *E. coli* strains expressing venom peptides at high levels will require a similar usage of both cysteine codons most likely to avoid depletion of one relatively to the other. Thus if one codon is present at higher frequencies then this will be more easily depleted within the cell and will become the limiting codon for rate of gene synthesis. Thus, the data suggest that codon usage is indeed one of the key determinants of expression yield. Regardless of the mechanism by which codon bias affects expression, systematic analysis of the relationship between gene sequences and expression will be a powerful tool to refine design algorithms, both for *E. coli* and other expression hosts.

Data presented here revealed that the best way to express high yields of folded active animal venom peptides is their expression in the periplasm of *E. coli*. In this study, DsbC was found to be a much more efficient fusion partner to express venom peptides in the periplasm of *E. coli* than MBP, which is widely-used at present (Bende *et al*., 2014, 2015; Cardoso *et al*., 2015; Klint *et al*., 2013, 2015; Saez *et al*., 2011; S. Yang *et al*., 2013). The encouraging results obtained with DsbC may derive from its excellent solubilisation potential but, more importantly, from its isomerase and chaperonin activities (which MBP lacks) that may promote the folding

of venom peptides. In the cytoplasmic compartment, the redox-inactive tags used (an inactive mutant derivative of DsbC and MBP) lead to the production of similar yields of recombinant peptides as the wild type DsbC. This suggests that redox properties of the fusion tag do not affect solubility and folding of animal venom peptides during the expression in the cytoplasm of *E. coli* and therefore confirm that oxidation occurs, primarily, *ex vivo* (Saez *et al*., 2014). During the production and purification of the proteins, DsbC is improving the solubility of venom peptides rather than assisting the peptides to reach their native oxidized form. Nevertheless, during the cleavage and in the presence of DTT, DsbC probably also acts as an isomerase as the yield of peptide varies greatly depending on the DTT concentration. Additionally, the peptide yields from the inactive mutant derivative of DsbC are much lower than the two other redox-active DsbC constructs. By using the DsbC fusion partner we were able to generate high yields of folded, putatively biologically active, venom peptides in the periplasm of *E. coli*. Data presented here, using TEV variant protein TEV_{SH}, a highly soluble TEV derivative, confirmed that all residues (except proline) could be accommodated in the P1´ position with little or no impact on the efficiency of processing, even in a non-optimized cleavage condition. These data confirm that for a project aiming at the production of a wide collection of recombinant venom peptides like the VENOMICS project, where individual optimization is impossible, removal of the associated fusion tag with TEV protease (using the shortened recognition site: ENLYFQ) can effectively produce a peptide with exactly the same sequence properties and biological activity as that of non-recombinant molecules.

There is an urgent need to develop effective methods to express large libraries of recombinant venom peptides that could be applied in innovative screening platforms for the discovery of novel therapeutics. *E. coli* is a highly robust heterologous host but it displays substantial limitations for the production of eukaryotic proteins with multiple disulfide bridges. Here we have analysed how to modulate the levels of expression, solubility and oxidation of animal venom peptides produced in bacteria. This report shows that under optimized conditions *E. coli* is a pertinent host for the expression of biologically active animal venom peptides. Genes encoding venom peptides and expressed at higher levels in *E. coli* present a codon usage that suggest a similar representation of the two Cys codons. This study demonstrates that the expression of venom peptides in the bacterial periplasm with the help of a DsbC fusion is one of the best options to purify milligram yields of active peptides, although the data suggest that peptide folding by DsbC occurs mainly *ex vivo*. Finally, with the exception of Pro, TEV protease can effectively tolerate any of the N-terminal amino acids located in venom peptides, suggesting that retention of the native peptide N-terminus is compatible with an effective protease cleavage. The application of the findings reported here helped us to build a high-throughput platform for the expression of venom peptides in *E. coli* to frame the VENOMICS project, leading to the production of the largest bank of animal venom peptides known (Chapter 6).

6. HIGH-THROUGHPUT EXPRESSION OF ANIMAL VENOM TOXINS IN *ESCHERICHIA COLI* TO GENERATE A LARGE LIBRARY OF RECOMBINANT RETICULATED PEPTIDES FOR DRUG DISCOVERY

Jeremy Turchetto[1*], Ana Filipa Sequeira[2,3*], Laurie Ramond[1*], Fanny Peysson[1*], Joana L.A. Brás[3], Natalie J. Saez[4], Yoan Duhoo[1], Marilyne Blémont[1], Catarina I.P.D. Guerreiro[3], Nicolas Gilles[5], Hervé Darbon[1], Carlos M.G.A. Fontes[2,3] and Renaud Vincentelli[1]

[1] Unité Mixte de Recherche (UMR) 7257, Centre National de la Recherche Scientifique (CNRS) Aix-Marseille Université, Architecture et Fonction des Macromolécules Biologiques (AFMB), Campus de Luminy, 163 Avenue de Luminy, 13288 Marseille CEDEX 09, France; [2] Centro Interdisciplinar de Investigação em Sanidade Animal (CIISA) - Faculdade de Medicina Veterinária, Universidade de Lisboa, Avenida da Universidade Técnica, 1300-477 Lisboa, Portugal; [3] NZYTech Genes & Enzymes, Campus do Lumiar, Estrada do Paço do Lumiar, Edifício E, r/c, 1649-038 Lisboa, Portugal; [4] Institute for Molecular Bioscience, The Universy of Queenland, St Lucia, 4072, Australia; [5] Commissariat à l'Energie Atomique, iBiTecS, Service d'Ingénierie Moléculaire des Protéines, 91191 Gif-sur-Yvette, France.

* Equal contribution

Adapted from a manuscript accepted in Microbial Cell Factories.

Abstract

Animal venoms are complex molecular cocktails containing a wide range of biologically active disulfide-reticulated peptides that target, with high selectivity and efficacy, a variety of membrane receptors. Disulfide-reticulated peptides have evolved to display improved specificity although, in contrast to proteins and antibodies, present low immunogenicity and show much higher resistance to degradation than linear peptides. These properties make venom peptides attractive candidates for drug development. However, recombinant expression of reticulated peptides containing disulfide bonds is challenging, especially when associated with the production of large libraries of bio-active molecules for drug screening. In the previous chapter an efficient system for the expression and purification of disulfide-reticulated venom peptides in *Escherichia coli* was described. Here we report the development of a high-throughput automated platform to generate the largest ever created library of recombinant venom peptides. The peptides were produced in the periplasm of *E. coli* using disulfide bond isomerase DsbC as a fusion tag, thus allowing the efficient formation of correctly folded disulfide bridges. Tobacco Etch Virus (TEV) protease was used to remove fusion tags and recover the animal toxins in the native state. Globally, from a total of 4992 synthetic genes encoding a representative diversity of venom peptides, a library containing 2736 recombinant disulfide-reticulated peptides was generated. The data revealed that the animal venom peptides produced in the bacterial host were natively folded and, thus, are putatively

biologically active. Overall this study reveals that high-throughput expression of animal venom peptides in *E. coli* can generate large libraries of recombinant disulfide-reticulated peptides of remarkable interest for drug discovery programs.

6.1. Introduction

Animal venoms contain a complex arsenal of disulfide-reticulated peptides that present an enormous structural and pharmacological diversity. The global animal venom library can be seen as a collection of more than 40,000,000 peptides of which only a very small fraction is known (Escoubas & King, 2009). These molecules have been fine-tuned during the course of evolution to present not only target selectivity but also low immunogenicity and high stability (Calvete, Sanz, Angulo, Lomonte, & Gutiérrez, 2009). The molecular targets of venom peptides are mainly present at the cell surface and are involved in various human health disorders such as pain, cancer, neurodegenerative diseases, cardio-vascular diseases, diabetes, obesity and depression (King, 2011; Lewis & Garcia, 2003). However, although the use of venoms for drug discovery is rapidly emerging it is still mostly an unrealized prospective due to recurrent technical bottlenecks, including the capacity to produce these biological highly relevant molecules recombinantly. For example, no comprehensive recombinant libraries of venom peptides are currently available for High-Throughput Screens (HTS) to identify novel therapeutics, as an alternative to synthetic chemical libraries.

Venom peptides generally contain between one to eight disulfide bonds which must be oxidized with the correct disulfide-bonding pattern in order to be active (Lavergne *et al*., 2015). Production of recombinant peptides in *Escherichia coli* has a number of advantages over other biological systems, including a reduced cost, rapid growth, high biomass production, easily scalable cultivation and well established regulations for therapeutic protein production. However, *a priori*, *E. coli* is not the ideal host to catalyse the formation of native disulfide bonds as its cytoplasm displays a particularly reducing environment (de Marco, 2009). Thus, proteins with disulfide bonds are especially prone to aggregation in *E. coli* as a result of mispaired intra- or inter-molecular disulfide bonds (Berkmen, 2012). In addition, production of recombinant proteins in bacteria is regulated by strong promoters, which favours the accumulation of misfolded recombinant proteins in the form of insoluble aggregates or inclusion bodies. Failure in reaching the bioactive conformation increases with the number of cysteine residues, due to the number of possible isoforms but also to the complexity of disulfide bond patterns. To overcome these disadvantages we developed a system for the efficient production of disulfide-bond-containing proteins and peptides in the cytoplasm of *E. coli* using a cleavable disulfide bond isomerase DsbC fusion in the strain BL21(DE3) pLysS (Nozach *et al*., 2013; and previous chapter of this thesis). The data revealed that although peptide folding by DsbC occurs mainly *ex vivo*, the expression of venom peptides when fused to DsbC in the bacterial periplasm lead to the production of milligram quantities of active toxins. The data also revealed that release

of fusion tags from recombinant peptides is effectively performed, even under non-optimal conditions by the use of Tobacco Etch Virus (TEV) protease.

The main goal of the VENOMICS project was to replicate *in vitro* the diversity of animal venoms in order to generate large peptide banks that could be applied in pharmacological screens used in drug discovery programs. Here a subset of about 5000 venom peptides with > 35 residues (as peptides with < 35 residues can be efficiently produced by chemical synthesis) representing the widest diversity in term of species, size, disulfide content and disulfide patterns was selected for the production of a recombinant animal venom peptide library. The results of the nine-months production phase indicated that out of the 4992 venom peptides selected for recombinant expression, 2736 (55%) peptides could be produced in a soluble and oxidized form in quantities compatible with the drug discovery program. Recently, this unique bank of recombinant venom peptides was screened to identify novel therapeutics for diseases such as diabetes, obesity, inflammation and allergies identifying dozens of novel drug leads. These data confirm that this unique library of animal venom peptides contains recombinant peptides that are both correctly folded and biologically active and represents an effective tool to discover novel drugs.

6.2. Materials and Methods

6.2.1. Gene synthesis and cloning

Genes encoding 4992 animal venom peptides originated from 201 different species, including terrestrials, marines and flying species, were designed for expression in *E. coli*. Table S6. 1 (see in Annex) describes the animal groups from which the peptides were selected. The number of genes to synthesise (4992) was selected considering the production of 52 plates in 96-well format (52×96 = 4992). Venom genes were designed by back-translating the peptide sequence and optimizing codon usage for high levels of expression in *E. coli*, using the ATGenium codon optimization algorithm. Global codon usage considers codons used preferentially in highly expressed or average native *E. coli* genes and exclusion of naturally rare codons, as well as ensuring an equal proportion of the two cysteine codons. GC content was set to vary between 40 and 60%. Presence of G/C islands, which could promote frame-shifting, was minimized by selectively avoiding runs of consecutive G and/or C greater than 6 nucleotides. In addition, no contiguous string of nucleotides longer than five nucleotides was allowed. Genes were designed to ensure the absence of *E. coli* regulatory sequences such as promoters, activators or operators. Codon Adaptation Index (CAI) estimates were calculated as the geometric mean for test gene codons of the ratio of the codon frequency in highly expressed *E. coli* genes divided by that of the highest frequency codon for each amino acid in those genes. Genes were designed to ensure a CAI value higher than 0.8. Genes were engineered to encode the TEV protease cleavage site (ENLYFQ/) at the N-terminus of the

target peptide sequence. This cleavage site was encoded by the sequence GAGAACCTGTACTTCCAA for all genes. The stop codon selected for all genes was TAA and was duplicated. The codon usage table used to design the DNA sequences of the 4992 genes is available in Table S6.2 (see in Annex). Synthetic genes were produced in a high-throughput pipeline using previously optimized procedures (Chapter 3). Briefly, genes were assembled from 40-60 bp oligonucleotides through PCR using KOD Hot Start DNA Polymerase (EMD Millipore). Oligonucleotides were designed using NZYOligo designer to have a maximum length of 60 bp and to ensure a 20 bp gap between primers located in the same strand. Oligonucleotides were synthesised by Integrated DNA Technologies at the smallest scale with desalting purification. PCR reactions were performed in a volume of 50 µL in 96-well PCR plates. After amplification, assembled PCR products were purified using NZYDNA Clean-up 96 well plate kit (NZYTech genes & enzymes, Portugal) in a Tecan workstation (Switzerland). Purified PCR products (~50 ng) were subsequently cloned into pHTP4 expression vector following established protocols for LIC technology (Chapter 3). In pHTP4 vector, venom peptides genes are under the control of a T7 promoter. Recombinant venom peptide fusion proteins contain an N-terminal DsbC fusion (with a signal sequence to export the protein to the *E. coli* periplasmic space), an internal six histidine tag for purification and a TEV recognition sequence to allow cleavage and isolation of the native peptide. Following the cloning reaction, recombinant plasmids were transformed using a high-throughput method into NZY5α competent *E. coli*. The transformed bacteria were spread on LB kanamycin agar plates. After overnight incubation at 37ºC, 1 colony per transformation was picked and grown in liquid media supplemented with 50 µg/mL of kanamycin in 24 deep-well-plates (5 mL) sealed with gas-permeable adhesive seals. The plasmids were purified from the bacterial pellets using NZYMiniprep 96 well plate kit (NZYTech genes & enzymes, Portugal) in a Tecan robot (Switzerland) and subsequently sequenced. In case the DNA sequence was not 100% identical to the designed gene, a second and eventually a third colony were picked for screening. All 4992 recombinant plasmids were completely sequenced in both directions to ensure consistency with the defined sequence.

6.2.2. High-throughput venom peptide preparation for drug discovery

Peptides were purified, characterized and prepared for functional test following a multi-step process in a 96 well plate format. First the DsbC-His-peptide fusions were purified from crude lysates using an automated nickel affinity procedure. The target peptides were further isolated on C18 resin after cleavage of the DsbC fusion partner by TEV protease. Correct mass, oxidation state and concentration of the resultant peptides were determined by liquid chromatography-mass spectrometry (LC-MS). Finally, the concentration of the oxidized venom peptides were adjusted to 10 µM (or 1 µM), aliquoted in multiples plates and frozen for the drug discovery process.

6.2.3. High-throughput venom peptide expression

All steps were carried out in 24 deep-well plates (DW24) with few modifications (see below) of the laboratory standard protocol (Saez *et al*., 2014; Saez & Vincentelli, 2014), which is briefly described below. 96 recombinant pHTP4 plasmids were used to transform BL21(DE3) pLysS *E. coli* cells in 96-well format, at a time. Transformed cells were used to inoculate pre-cultures in 96 deep-well (DW96) plates containing 1 mL of LB medium in each well. 50 µL of the pre-culture broth (1/40 v/v) was used to inoculate 2 mL of auto-induction medium. Cultivation was carried out using DW24 plates. For the production in each DW24 plate, 12×2 mL of auto-induction medium supplemented with kanamycin (50 µg/mL) and chloramphenicol (34 µg/mL) was used for each peptide. Downscaling the culture volume from 4 mL to 2 mL, with a better aeration, doubled the production yield compared to the previous protocol. Since the aim of the VENOMICS project is to produce a comprehensive library of peptides in HTS format (250 µL at 10 µM) and taking into account the final average peptide yield obtained with the periplasmic DsbC fusion (described in the previous chapter) a scale of 12x2 mL culture/toxin (24 mL) seemed to be the best compromise between yields, labour time and cost. To express 96 peptides in parallel, a series of 48×DW24 were inoculated simultaneously (1152 cultures) and grown over 24 h at 25ºC in a Microtron shaking incubator (INFORS-HT, Switzerland) at 600 rpm. The protein expression at 25ºC was preferred as it gave slightly higher peptide recovery yield than the previous procedure (37ºC for 4 hours followed by 18h at 17ºC). To be able to put simultaneously 2×24×DW24 in the incubator simultaneously, a plexiglass second level was custom-made and added inside the incubator. The ZYP-5052 medium was replaced by the NZY Auto-Induction LB medium (NZYTech, genes & enzymes, Portugal) to gain efficacy, without any significant difference in protein yield (data not shown). At the end of the culture, the OD_{600nm} was 12, on average. Cells were collected by centrifugation and each well re-suspended in 0.5 mL of lysis buffer (50 mM Tris, 300 mM NaCl, 10 mM Imidazole, pH 8.0, 0.25 mg/mL lysozyme) by 15 minutes shaking at 20ºC in a Microtron shaking incubator (INFORS-HT, Switzerland) at 800 rpm. The DW24 plates were then frozen at -20ºC.

6.2.4. High-throughput protein purification by nickel affinity chromatography

12×DW24 plates (corresponding to the over-expression of 24 toxins) were thawed in a water-bath for 10 minutes at 37ºC, followed by 15 minutes shaking at 20ºC in a Microtron shaking incubator (INFORS-HT, Switzerland) at 800 rpm. During this time the cultures were lysed by lysozyme and the solution became viscous. After the addition of 10 µg/mL DNAse and 20 mM $MgSO_4$ in each well and incubation for 10 minutes at 20ºC in a Microtron shaking incubator (INFORS-HT, Switzerland) at 800 rpm, the 12x0.5 mL lysed bacteria for each peptide were pooled and transferred to a new DW24 plate to have 24 distinct 6 mL lysates of DsbC-His-peptide fusions per DW24 plate. To ensure complete cell lysis, the DW24 was sonicated in a plate sonicator (Ultrasonic processor XL, Misonix Inc., USA) for 5 minutes (power 5, 30

seconds ON/OFF cycles). Purification was performed by Immobilized Metal Affinity Chromatography (IMAC) as described in the previous chapter where all the steps were automated on a Tecan Freedom EVO 200 robot (Switzerland) containing a vacuum manifold. Briefly, the 6 mL of crude cell lysates were incubated with 4×200 µL Ni^{2+} Sepharose 6 Fast flow resin (GE Healthcare) with bound Nickel and then transferred (4 wells per peptide) into 96-well filter plates with 20 µm (Macherey-Nagel). The wells were washed twice with 800 µL of buffer A (50 mM Tris, 300 mM NaCl, 10 mM Imidazole, pH 8.0) followed by three washes with 800 µL of buffer B (50 mM Tris, 300 mM NaCl, 50 mM Imidazole, pH 8.0). The recombinant fusion proteins were eluted from the resin beads with 1 mL of elution buffer (50 mM Tris, 300 mM NaCl, 250 mM Imidazole, pH 8.0). Rather than collecting the elution in a DW96 plate, like in the standard procedure, this DW plate was replaced by a DW24 plate, pooling the 4×1 mL of elution of a single peptide into a single well of the DW24. This procedure was reproduced four times in one day in order to purify the 96 fusion proteins. The total time taken for a single round of this process was around 4 hours, however, the end of one round can be performed concurrently with the beginning of the next round.

6.2.5. High-throughput TEV cleavage

On the same day, after the 96 purifications, the concentration of the 96 purified DsbC-His-fusion proteins was calculated spectrophotometrically (OD at 280 nm) in a micro-titre plate reader (Genius plus, TECAN, Switzerland). A mutant of TEV protease (TEV_{SH}) (Van Den Berg *et al*., 2006) (stored at 2 mg/mL in 20 mM Hepes, 300 mM NaCl ,10% Glycerol, pH7.4 buffer; see previous chapter for the production protocol) was added (1/10 w/w) directly in the elution buffer by the Tecan robot. The final concentration of DTT was adjusted to 0.1 mM with a fresh DTT solution while adjusting the final cleavage reaction volume to 5 mL with elution buffer. The 96 samples were then incubated overnight at 30ºC in a Microtron shaking incubator (INFORS-HT, Switzerland) at 200 rpm to allow total cleavage of the DsbC-His-fusion protein.

6.2.6. High-throughput target peptide purification by reverse phase chromatography

After overnight TEV cleavage, samples were acidified for 1h with 5% acetonitrile, 0.1% formic acid (FA) by adding 500 µL of a 55% acetonitrile and 1.1% formic acid stock solution, at room temperature under mild agitation. At this step, TEV protease and mis-folded peptides were precipitated by the acidification and the 96 samples were subjected to the final purification step on C18 chromatographic resin. The C18 purification was performed using an automated Solid Phase Extraction (SPE) procedure specifically developed for this project and was implemented on a Tecan robot. First, the 4×DW24 containing the 96 acidified samples were centrifuged at room temperature (4000 ×*g* for 10 minutes) and the supernatants (5 mL) were transferred (5×1 mL with application of vacuum between each cycle to remove the unbound materials) into a

96-well filter plate with 20 µm (Macherey-Nagel) filled with 50 µL of C18 reversed phase beads (100 Å, Sigma) that had been activated in pure acetonitrile and equilibrated in solvent A (5% acetonitrile, 0,1% FA). After loading the samples, the C18 resin was washed twice with 800 µL of solvent A followed by one wash of 800 µL of solvent A without FA. The cleaved DsbC fusion partner is eliminated in the flow through and wash fractions of the SPE purification. The pure target peptides were eluted from the resin beads in 560 µL of elution buffer (50% acetonitrile/water). At the end of the SPE, the DW96 containing the target peptides in 560 µL of elution buffer (50% acetonitrile/water) were maintained on mild agitation under a chemical hood during 16 hours so that the acetonitrile evaporated. Thus, the peptides were obtained in 280 µL of pure water. The peptides were stable for several weeks at 4ºC in these conditions (data not shown).

6.2.7. High-throughput quality control and quantification by mass spectrometry

A first aliquot of 10 µL was taken from the 280 µL of purified target peptides to have an external confirmation than the peptide oxidation states were correct (University de Liege, Belgium). A second aliquot (20 µL) of the 96 cleaved samples was analysed in-house on a reverse phase C18 column (Hypersil GOLD 50 x 1.0 mm, 1.9 µm, 175 Å, ThermoScientific) at 37ºC, with at a flow rate of 200 µL/min on an Ultra-high performance liquid chromatography-mass spectrometry (UHPLC-MS) with electrospray ionization (Accela High Speed LC system with detector MSQ+, ThermoScientific, San Jose, California). The gradient slope (solvent A: water, solvent B: acetonitrile, both solvents containing 0.1% formic acid) went from 5 to 40% B in 2 min followed by an 80% wash and re-equilibration (total time: 6 minutes). MS acquisition was performed in the positive ion mode from *m/z* 100 to 2000. To confirm correct target peptide molecular weight, the resulting mass spectra were de-convoluted using manual calculations. The isotopic pattern measured was compared with the theoretical one determined from the amino acid sequences using Data Explorer software (version 4.9, Applied Biosystems). The quantitative calculation of target peptide yields was determined using automatic processing with Xcalibur™ software (ThermoScientific), by OD_{280nm} measurement and peak areas integration.

6.2.8. Venom peptide bank preparation for high-throughput screening

Since the final peptide quantification method was based on the $OD_{280\ nm}$ measurement and integration of peak mass on the UHPLC, and because we estimated that this quantification mode (the only one compatible with the throughput of the process) could lead to errors of up to 100% on the concentration for peptides with low molar extinction coefficients, we decided to divide the peptides into three subsets, based on concentration, for the screening process. When the peptide concentration was above 20 µM, the quantification was considered accurate and the peptide concentration was adjusted to 10 µM before aliquoting. When the

concentration ranged from 5 to 20 µM, the peptides were aliquoted at this concentration. When the peptide concentration ranged from 1 to 5 µM, the peptides were aliquoted as such in a separate series of plates (see below). When the concentration was below 1 µM, the peptides were discarded and counted as a non-producing clone. From this information, the most concentrated toxins (concentration > 20 µM) were diluted with water, using the Tecan robot, to a final concentration of 10 µM (v=250 µL) in a DW96 plate. Each DW96 plate contained 80 target peptides and 2x8 empty wells for assay controls. From this original DW96 stock, the Tecan robot was used then to made 5 copies of microtiter plates containing 50 µL of each target peptide. These plates are named the “10 µM bank”. The plates were shock frozen and stored at -80ºC until delivery to the screening laboratory (the delivery was done in 2 batches). Remaining concentrated peptide was shock frozen directly in the DW96 and kept as a backup for validation of potential hits, post-screening. When the peptide concentration ranged between 5 and 20 µM, the 250 µL left after the evaporation of the SPE fractions were reorganized in DW96 and treated as above to generate 5x50 µL plates. These plates were also called the “10 µM bank”. Finally, when the concentration was ranging from 1 to 5 µM, the peptides were reorganized in DW96 and treated as above to generate 5x50 µL plates. These plates were called the “1 µM bank”. The acquisition of the Liquid Chromatography-Mass Spectrometry (LC-MS) data, the analysis of the spectra, the Tecan concentration adjustment and aliquoting phases took less than 24 hours for 96 peptides.

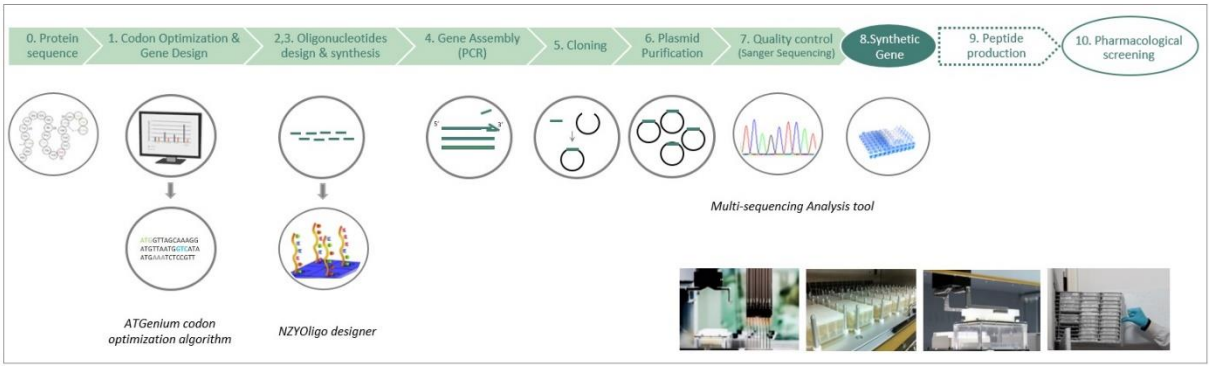
6.3. Results

The time frame allocated within the VENOMICS project to establish the largest possible recombinant, active, disulfide-reticulated venom peptide bank was 9 months. The high-throughput pipeline and the results depicted below were implemented to cope with this time constraint.

6.3.1. Generation of a library of *E. coli* expressing plasmids encoding 4992 venom peptides

A previously optimized high-throughput gene synthesis platform (see Chapter 3) was used to synthesise 4992 genes encoding animal venom peptides. The platform is automated for the majority of its steps using 24 or 96 -well plates, a standard high-throughput liquid-handling Tecan robot and various bioinformatics tools. A schematic representation of the platform is presented in Figure 6.1.

Figure 6.1| HTP gene synthesis platform used to produce 4992 synthetic genes encoding venom peptides.



This pipeline includes 7 steps that allow the successful synthesis of multiples of 96 genes. The first step corresponds to gene design and codon optimization; from multiple peptide sequences, DNA sequences are designed and optimized for expression in *E. coli*, using the ATGenium codon optimization algorithm. In steps 2, 3 and 4 oligonucleotides required for gene assembly are designed using the NZYOligo designer, synthesised and assembled by PCR using optimal conditions, respectively. Synthetic genes are cloned using NZYTech LIC protocol into the *E. coli* expression vector pHTP4. Bacterial transformation and DNA preparations are accomplished using high-throughput protocols. DNA sequences are checked for the presence of sequence errors using a high-throughput Sequencing Analysis tool. All steps are automated using a Tecan Freedom EVO 200 robot (Switzerland).

The gene synthesis process was performed according to the timeline of VENOMICS project, with the production of 4×96 synthetic genes per week. The primary sequences of 4992 reticulated peptides originated from 201 venomous animal species were used to design genes for optimal expression in *E. coli*. The algorithm was set to lead to a similar incorporation of the two cysteine codons, as suggested by data presented in the previous chapter. Genes contained an average GC content of 49% and an average CAI of 0.92 (Table 6.1). Oligonucleotides used to assemble the gene library were designed to have an overlap region of 20 bp and a gap of 20 bp, while having a maximum length of 60 bp. In average 6 primers were required to assemble each nucleic acid and genes had an average size of 220 bp (Table 6.1).

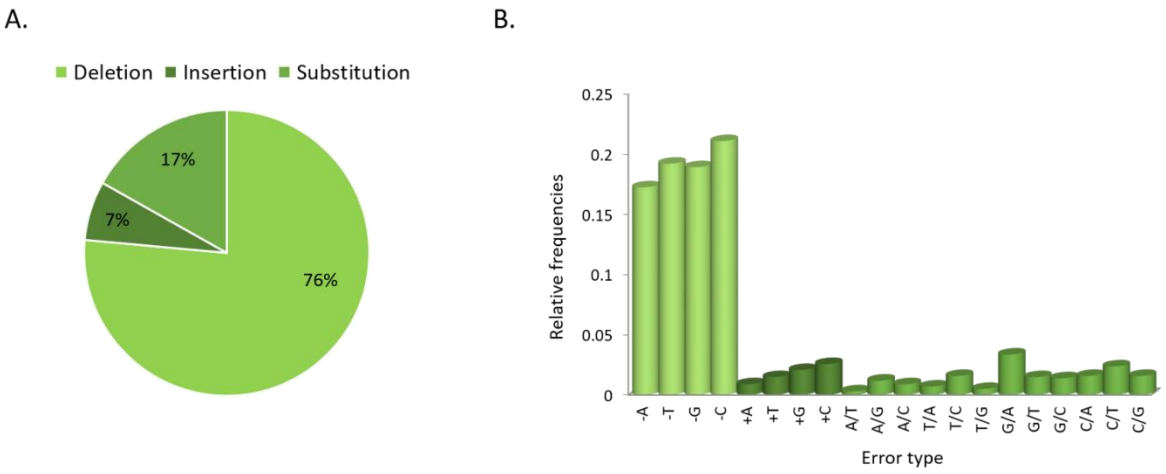
Table 6.1| Properties of the 4992 genes synthesised in this project.

	Length (bp)	GC content (%)	Number of primers	Codon Adaptation Index (CAI)
Mean (± SD)	220 ± 54	49 ± 4	6 ± 1.6	0.92 ± 0.04
Maximum	413	58	10	0.94
Minimum	137	42	4	0.8

Individual genes were PCR assembled and after nucleic acid purification directly subcloned into pHTP4 expression vector using a LIC method. No methodology was implemented to correct eventual mistakes arising during gene synthesis. The robustness of the pipeline was

demonstrated when plasmid DNA was sequenced to verify gene integrity. After screening only one colony per cloning reaction, 3818 of the genes (76.5%) were observed to be correct, meaning that for more than 75% of the gene synthesis reactions it was possible to easily recover a gene that accumulated no mutations during nucleic acid assembly. Thus, a second colony was inoculated for 1174 LIC reactions for which no correct clones were obtained in the first screen. When the second colony was screened, 809 of the genes (16.2% of the total of genes) were found to be correct. Finally, for 365 genes (7.3% of the total of genes) it was necessary to pick a third colony to obtain a correct nucleic acid. Taken together the data revealed that it was necessary to screen an average of 1.3 clones to obtain the correct gene. Therefore, this gene synthesis platform exhibited an error rate of 1.06 errors/kb. The majority of the detected errors were deletions (76%), as is expected from the oligonucleotide synthesis methodology (LeProust *et al*., 2010). Insertions were less common (7%) while substitutions represent 17% of the identified errors (Figure 6.2). The most frequent substitution identified in incorrect genes was a G for an A. Deletions or insertions of a C were the most represented ones while the nucleotide A was less prone to be erroneously inserted or deleted (Figure 6.2).

Figure 6.2| Errors observed during the synthesis of 4992 genes encoding venom peptides.



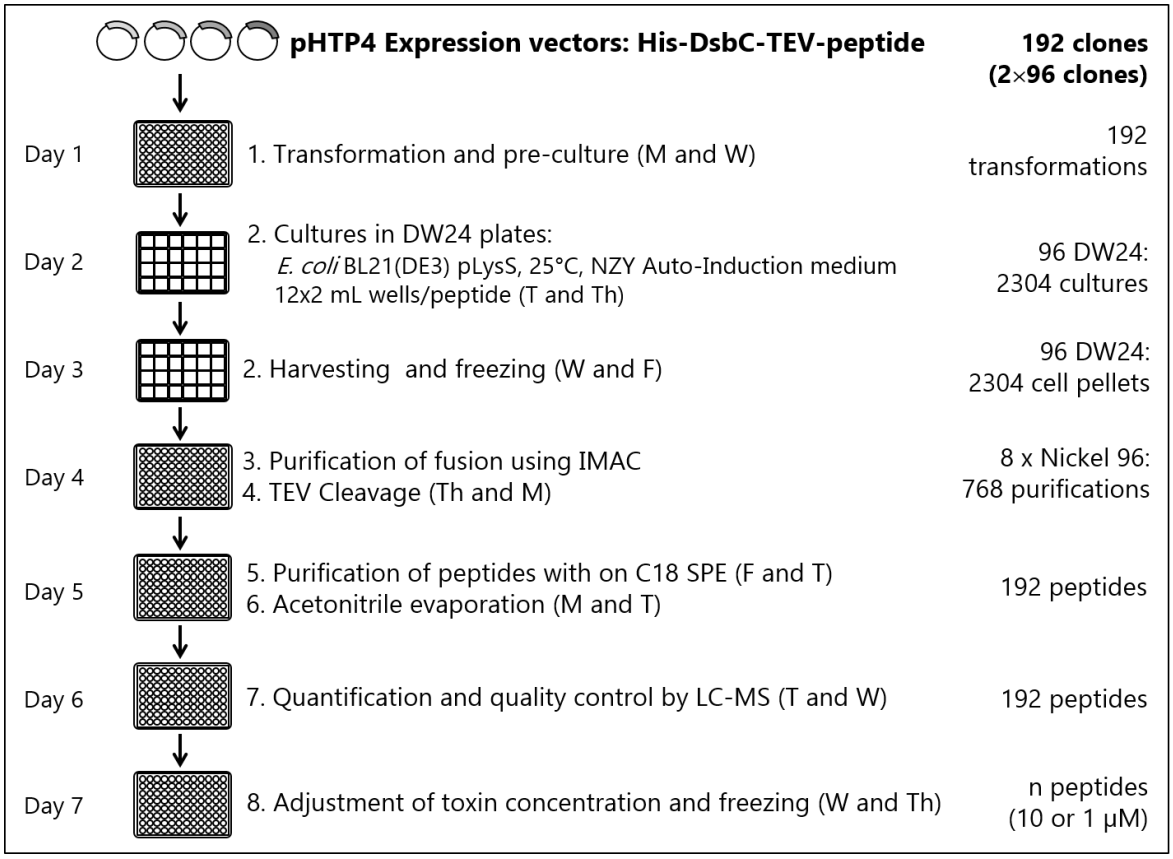
In Panel A, the percentage of types of errors identified in the genes is described. In Panel B, the type of mistakes that were observed are specified.

6.3.2. Generation of a library of 2736 oxidized venom peptides for drug discovery

After optimizing several parameters for the production of oxidized disulfide-reticulated peptides in *E. coli* (see Chapter 5), a new high-throughput protocol was assembled to try to produce, from a library of 4992 plasmids, a maximum number of oxidized recombinant target peptides in a nine-month period. The platform was mostly automated on a Tecan Freedom Evo liquid-handling robot (Switzerland) using 24 or 96 -well plates. The process is made of roughly eight steps. A schematic representation of the pipeline used for the recombinant production of

animal venom peptides in *E. coli* is presented in Figure 6.3. To cope with the timeline of the project, the purification pipeline (96 peptides) was run twice every week, so the team of three people performed 2304 cultures, 768 nickel purifications, 188 SPE and 188 LC-MS every single week. At the end of the production (9 months), each peptide had been through the production pipeline only once. It was successfully implemented a complex process made of eight steps taking from seven to eight days (from the initial culture transformation to the frozen normalized peptide bank; see Figure 6.3). This effort represents, altogether, more than 60,000 single cultures, 5000 affinity purifications and 5000 reverse phase purification followed by 5000 LC-MS quality controls and quantification.

Figure 6.3| Schematic representation of the high-throughput pipeline used for the production of recombinant venom peptides in *E. coli*.



Week schedule used to produce 2×96 recombinant toxins per week. Plates numbered with even numbers started on Mondays and plates odd numbered started on Wednesdays. Between bracket are the days of the week (M, Monday; T, Tuesday; W, Wednesday; Th, Thursday; F, Friday).

From the 4992 peptides that entered the pipeline, 4963 (99.4%) were analysed by LC-MS. The 29/4992 peptides (0.6%) that were lost before the end of the production, failed to give viable colonies or cultures. Due to time constraints these cultures were not reproduced. The cultivable 4963 peptides (99.4%) were successfully purified by IMAC. Among these, 66 peptides (1.3%) were initially lost during the purification (clogged wells or cross contamination due to partial flooding of a neighbouring well). These 66 peptides were combined in a salvage experiment,

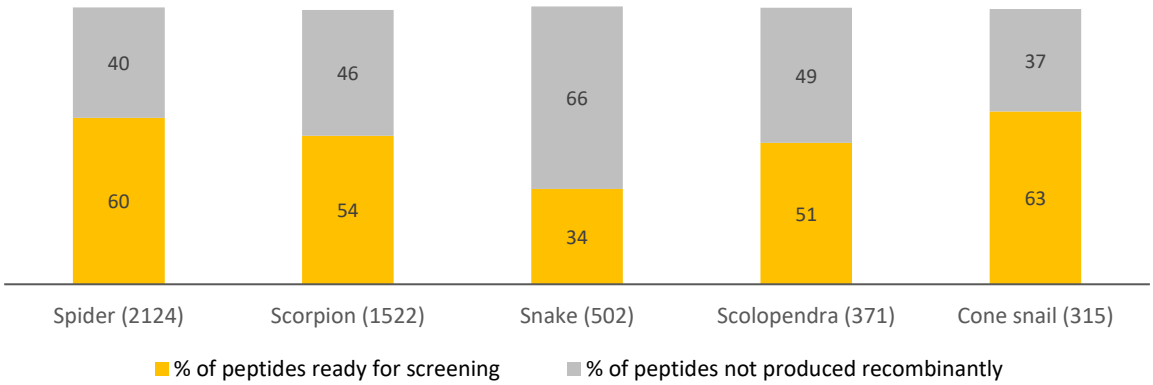
the cultures were reproduced and the 66 peptides reached the LC-MS step on a second batch. Thus, from the 4992 peptides that entered the pipeline, 4963 went up to the affinity purification step and 2736 (55%) were produced with the correct peptide molecular weights, corresponding to their fully oxidized isoforms. The characteristics of each peptide and the output of their production are presented in Table S6.3 (see Annex). From the 2736 peptides of the bank, 2174 (~80%) had a concentration above 5 µM and were aliquoted in the “10 µM bank” (28 plates), while the remaining 562 (~20%), with a concentration ranging from 1 to 5 µM, were aliquoted in the “1 µM bank” (8 plates). Among the 10 µM bank, 1363 (~ 63%) were purified at a concentration above 20 µM. The resulting bank contains to date, the biggest and most diverse collection of recombinant animal venom peptides available for drug discovery screening. These results demonstrate the extreme robustness of the production and purification pipeline. Out of the 4963 purified fusion peptides, 4843 (97.6%) DsbC-His-peptides could be purified in milligram quantities (per litre of culture) after IMAC purification. In theory, with a 100% cleavage efficiency and recovery, these should have allowed the purification of peptides in the levels required for the drug discovery pipeline. The average yield from these 4843 recombinant strains was 183 milligram of purified fusion protein per litre of culture with a maximum yield of around 500 mg/L. These data are in accordance with the result obtained in the previous chapter. Thus, the average yield per litre reflects the purification of a total of 4.4 mg of fusion peptide using the pipeline (from 24 mL culture). However, to avoid compromising the throughput of the pipeline, yields of purified fusion peptides were not determined on Caliper or LC-MS, which could have provided information about presence of truncated derivatives of the proteins. Concentration of the purified protein was quantified only by OD at 280 nm. These concentrations are therefore probably overestimated for a portion of the population of the DsbC-His-peptides. Indeed, in the previous chapter we identified, by Caliper analysis of the purified fusions, that in some cases the correct population (DsbC-His-peptide) was contaminated by truncated forms of the fusion protein (DsbC-His alone). Notwithstanding these observations, from the 2736 peptides aliquoted for drug discovery (with an average mass of 6127 Da that compares with an average mass of 6424 Da for the full bank), the average concentration of purified recombinant peptide obtained at the end of the last purification step was 58.5 µM (in 250 µL), leading to an average yield of 3.91 mg of peptide per litre of culture (95 µg/15 nmoles in the 250 µL, 0.38 mg/mL concentration, from the 24 mL culture scale). Thus, the average recovery of toxins (11.3%), is low and below the values found in the previous chapter of these thesis. This could be explained by several reasons. First, as explained above, the yield of purified fusion is in some cases over-estimated. Secondly, for some peptide families that where not present in the previous tests (see previous chapter), the TEV cleavage might not be complete due to non-optimal cleavage conditions (TEV ratio, DTT concentration in the buffer, etc.). In addition, following the present protocol, peptides are purified from the DsbC-His fusion by a C18 purification step that was not present in the pipeline described in

the previous chapter. Finally, peptides with non-native inter-chain bonds do precipitate in this rather harsh protocol (all steps done at room temperature, quick acidification, etc.). This is confirmed by the fact that the non-correctly oxidized or mixed forms (intermolecular disulfide bridges) peptides were never detected on the LC-MS. For a smaller project these yields could be increased by carefully checking the quality of the proteins at each step and by re-screening TEV cleavage conditions.

6.3.3. The pipeline is efficient for the production of venom peptides independently of animal origin, peptide length, cysteine patterns or number of disulfide bridges

One of the objectives of this study was to set up a collection of peptides representing the natural diversity in animal venom. The panel selected here was issued from a wide variety of animals with variable peptide length (35 to 120 amino acids), number of disulfide bridges (1 to 9) and cysteine patterns (84 different cysteine patterns). The 4992 peptides that were selected for the recombinant production were originally identified from 201 animal species. The most represented species were spiders (2124 samples, 43%), scorpions (31%), snakes (10%), centipedes (scolopendra) (7%) and cone snails (6%). Several other species were also present (ants, anemones and various insects, 3% altogether, with an average success rate of 60%). While peptides sized 35 amino acids or longer were produced recombinantly, within VENOMICS smaller peptides were produced via chemical synthesis (performed by other VENOMICS partner). The chemistry main effort was focused on the synthesis of cone snail (56%), spider (29%) and scorpion (8%) peptides. The success rate of the production of recombinant animal venom peptides origin is displayed in Figure 6.4. Overall, the data suggest that with the exception of peptides from snakes, with a success rate of only 34%, the source origin had little impact on the capacity to obtain peptides recombinantly.

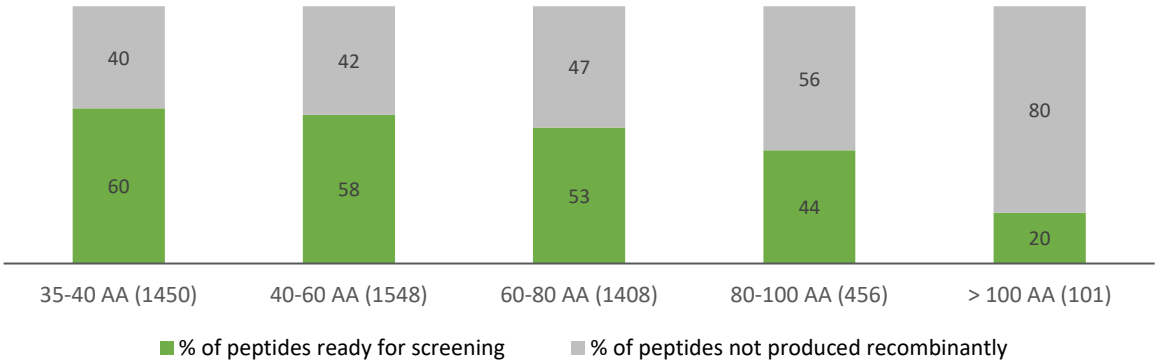
Figure 6.4| Effect of venom peptide origin in the success rate of production.



The number of peptides analysed for the different animals is presented in brackets. In gray: percentage of peptides not produced successfully. In yellow: percentage of animal peptides produced in sufficient quantities for screening.

The length of peptides produced within this study span from 35 to 120 amino acids. Data presented in Figure 6.5 suggest that although the pipeline was successfully expressed the majority of peptides, the success rate decreased with increasing peptide length. The data confirm a drastic decrease in the capacity of producing recombinant peptides when sizes are bigger than 100 residues. However, these large size peptides were underrepresented in the 4992 bank as they correspond to 2% of the plasmid bank.

Figure 6.5| Effect of peptide length in the success rate of production.



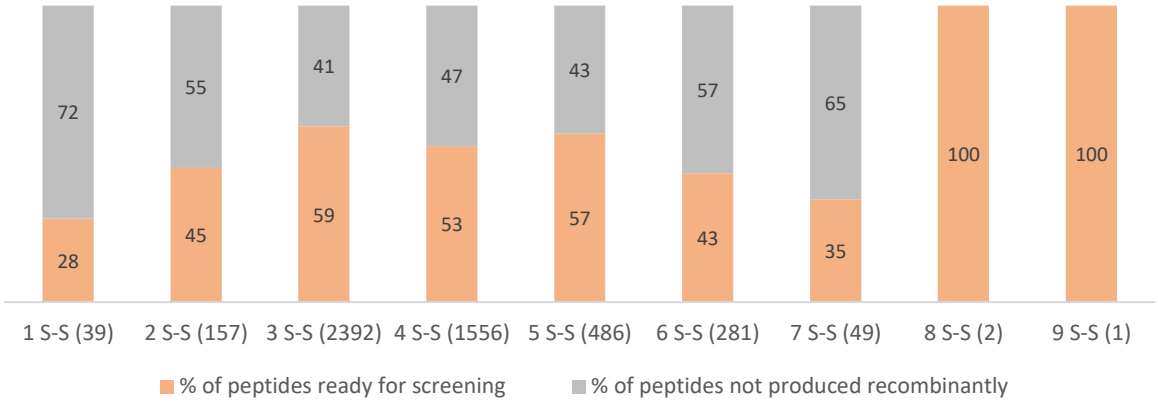
The number of peptides analysed for the different classes of peptide length is presented in brackets. In gray: percentage of peptides not produced successfully. In green: percentage of peptides produced in sufficient quantities for screening.

Within polypeptides, cysteine patterns reflect the number and distribution of cysteines along the primary sequence. From the 84 cysteine patterns analysed in this work, the majority were successfully produced as correctly folded toxins (67/84, see Figure S6. 1 in Annex). Except for 17 patterns (representing 0.6% of the targets) where no oxidized peptide could be detected, the other patterns exhibited some success demonstrating the rather wide spectrum capacity offered by the DsbC-fusion expression system. These patterns span from the shortest “C-C” peptides (one disulfide bridge, 39 peptides; 28% successful recovery), to peptides with 8 and 9 disulfide bridges (that could be fully recovered). Some patterns that are highly represented could be tentatively associated with putative disulfide bridge reticulation, and therefore with peptide 3D structures. For example, C-C-CC-C-C (1221 occurrences, 67% success rate, see Figure S6. 1 in Annex) is the most abundant cysteine pattern found in this study and it can easily be associated with the inhibitor cystine knot (ICK) motif, highly present in spider venom peptides. While only one has been described in the past for scorpions, this study includes 15 non-redundant new scorpion sequences. Three fingers snake peptides can be recognized in the following pattern: C-C-C-C-C-CC-C (471 occurrences). Finally, C-C-C-C (122 occurrences, 53% success) could well stabilize the secondary structure found in anemone peptides and was rarely described for scorpion venom peptides to date. The pipeline was more efficient at producing peptides with 3, 4, 5 disulfide bridges, which represent 70% of the peptide bank

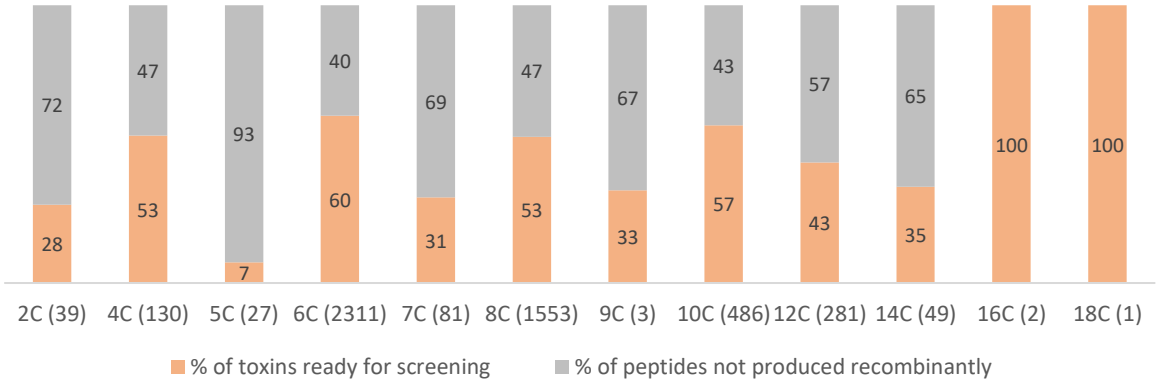
(Figure 6.6, A). The success rate slightly decreases for more complex peptides (6 and 7 bridges). Remarkably, the three most complex peptides, containing 8 (2 peptides) and 9 (1 peptide) disulfide bridges, were successfully produced while the pipeline turned out to be comparatively inefficient for peptides with reduced number of disulfide bridges. Only 28% of peptides containing only one disulfide bridge were obtained in the recombinant form. A detailed analysis of the 4992 primary sequences comprising the peptide bank revealed the presence of 111 peptides with an odd number of cysteines, a feature not common in the venom peptide field (Lavergne *et al*., 2015).

Figure 6.6| Effect of number of disulfide bridges (Panel A) and number of cysteine residues (Panel B) in the success rate of production.

A.



B.



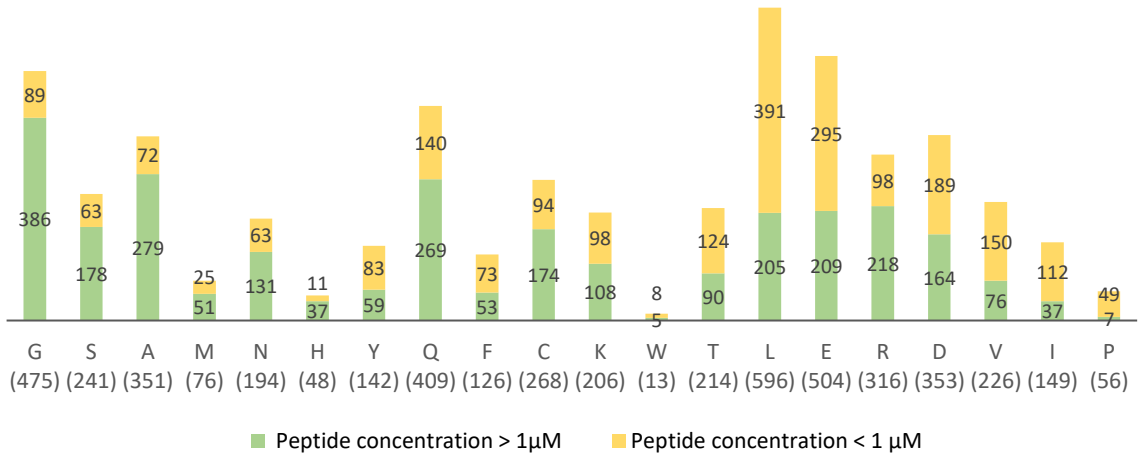
The number of peptides analysed in the two different cases is presented in brackets. In gray: percentage of peptides not produced successfully. In orange: percentage of animal toxins produced in sufficient quantities for screening.

Data, presented in Figure 6.6, B, revealed that for peptides containing an odd number of cysteines, the success rate of production was much lower when compared with peptides containing a paired number of cysteines. This observation is intriguing and more work is

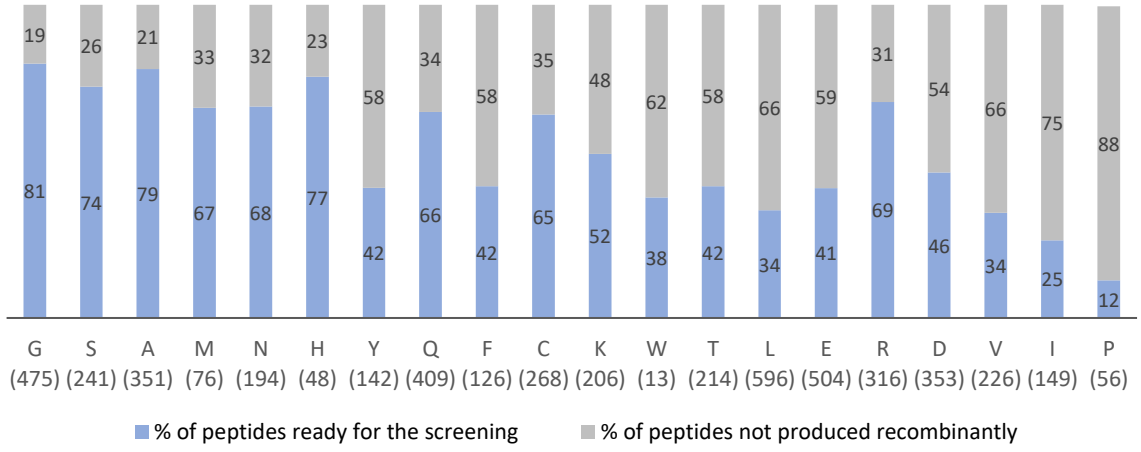
required to confirm whether these selected sequences resulted from annotation or curation errors rather than new kinds of venom peptides. The N-terminal residues of venom peptides can contribute to the ligand binding site. It is, therefore, possible that introduction of a single, extra residue, at the N-terminus of recombinant peptides may affect their biological activity (Karbat *et al*., 2007). Data presented in the previous chapter suggested that, with exception of proline, all residues may be accommodated at the P1´ position of the TEV protease recognition site without notably affecting the efficacy of processing. Thus, in this study, no extra residue was engineered at the N-terminus of the recombinant peptides to improve TEV cleavage efficacy. Analysis of the primary sequences of the 4992 peptides revealed that all 20 amino acids can be found at the N-terminus of the native peptide sequence (Figure 6.7).

Figure 6.7| The nature of the N-terminal residue in native venom peptides affects the success rate of production.

A.



B.



Panel A displays the number of peptides containing the 20 different amino acids at the N-terminus. In green: number of peptides produced at > 1µM. In yellow: number of peptides with concentration < 1 µM. In Panel B, the percentage of peptides produced or not produced is displayed. The number of peptides analysed is presented in brackets. In gray: percentage of peptides not produced successfully. In blue: percentage of peptides produced in sufficient quantities for screening.

The most represented N-terminal amino acids in venom peptides of the 4992 sequences are leucine, glutamate and glycine (around 10% of the bank for each one) (Figure 6.7, A). N-terminal cysteines, which should also be involved in disulfide-bounding, are relatively abundant in the population while peptides starting with a proline, histidine or tryptophan are very rare (1% or less). The relation between nature of the N-terminal residue and capacity to produce the corresponding peptide was analysed. The data, presented in Figure 6.7, B, suggested a similar trend to cleavage efficiency when compared with data presented in the previous chapter. Thus, this observation indicates that the folding of the peptide had little impact on the cleavage efficiency, which is more affected by the nature of the N-terminal residue of the peptide. Strikingly, cleavage was very effective for peptides containing a cysteine in P1' position, suggesting that presence of residues participating in disulfide bridges at the N-terminus of the polypeptide does not interfere with the TEV protease access. In addition, even if uncommon (12.5%), it was possible to cleave off some peptides containing a proline at P1' position. Thus, it is possible that in particular proteins the primary sequence downstream the proline residue allows it to assume a conformation that is more permissible for proteolytic attack.

6.4. Discussion

There is an urgent need to develop a novel model for modern drug discovery. Monoclonal antibodies have recently become increasingly attractive therapeutics, delivering important results in the treatment of several major disorders including autoimmune disease, cancer, inflammation, cardiovascular and infectious diseases. The fact that antibodies have become captivating therapeutic agents motivated a recent interest in other biological molecules, and particularly peptides, as valid leads for innovative drugs (Craik, Fairlie, Liras, & Price, 2013). Animal venoms constitute a vast and essentially untapped resource of novel biologically active molecules and thus should play a prime role in modern drug discovery (Sébasti Dutertre *et al*., 2015; Harvey, 2014; Takacs & York, 2014). Venom evolved in a wide variety of invertebrates (e.g., sea anemones, corals, jellyfish, marine molluscs, spiders, scorpions, hymenopteran insects and marine worms), as well as vertebrates, such as snakes (Escoubas *et al*., 2008). It is now well established that venom peptide targets are involved in various human pathologies such as pain, cancer, neurodegenerative diseases, cardio-vascular diseases, diabetes, obesity and depression. Thus, animal venoms constitute vast libraries of pharmacologically active disulfide-reticulated peptides, which have evolved to be highly selective receptor-targeted molecules with low immunogenicity and high stability (Lewis & Garcia, 2003). However, replicating nature´s venomic diversity *in vitro*, to generate large collections of bioactive reticulated peptides that could be pharmacologically screened to identify novel drug leads is still, mostly, an unrealized prospect. Genuinely, the recombinant expression and

refolding of reticulated peptides has been a major limiting factor in the use of venoms for drug discovery.

With the aim of exploring the huge biodiversity presented by animal venoms, we have developed and optimized a high-throughput pipeline for the competent production of large libraries of bioactive recombinant venom peptides. Data presented here reveals that *Escherichia coli* is an effective heterologous host to express large numbers of recombinant fully-oxidized venom peptides. The genes encoding around 5,000 venom peptides from different sources were *de novo* synthesised with a codon usage optimized for expression in *E. coli*. A total of 1100 kb of DNA was synthesised during this project. The gene synthesis method employed here exhibited a low error rate of 1,06 errors/kb, which is similar or slightly better when compared with other gene synthesis methodologies based on PCR or ligation assembly (Ma *et al*., 2012; Saaem *et al*., 2012; Schwartz, Lee, & Shendure, 2012; Wan *et al*., 2014). As observed here, regardless of the synthesis method, incorrect bases are likely to be incorporated into the gene sequence during assembly. This arises mostly due to errors accumulated during the chemical synthesis of the oligonucleotides (LeProust *et al*., 2010). Thus, the most common errors observed during *de novo* gene synthesis are deletions that result from the incorporation of truncated versions of the oligonucleotides used for gene assembly. Typically, these errors require removal to retrieve the integral DNA sequence (Ma *et al*., 2012; Xiong *et al*., 2008), although this process introduces complexity in the gene synthesis process, which most of the time is not compatible with high-throughput methods. Various strategies were developed to reduce encoded errors, including use of enzymatic mismatch cleavage (Fuhrmann *et al*., 2005; Saaem *et al*., 2012), mismatch-binding proteins (Carr & Church, 2009) and site-directed mutagenesis (Xiong *et al*., 2008). Since genes encoding venom peptides are relatively small (average size of 220 bp) a low error rate was expected and, therefore, an error correction technology was not employed in this pipeline.

Here, a high-throughput platform was used to establish a large library of recombinant venom peptides for drug discovery. The library contains 2736 different oxidized peptides of different origins validating the capacity of *E. coli* to produce bio-active disulfide-reticulated peptides. The major factor affecting the capacity of bacterial cells to express animal venom peptides is peptide length, as the efficiency of production dramatically decreases for peptides of more than 100 amino acids. Overall the percentage of correctly oxidized peptide is much reduced when compared with the production of the DsbC-His-fusion peptide derivative, suggesting that a considerable fraction of the peptides is not properly folded when produced in the fusion form. Nevertheless, using the DsbC fusion partner and directing the expression to the periplasm allows significant levels of biologically active venom peptides to be obtained in *E. coli*. In addition, it is possible that a significant number of peptides were not properly produced due to difficulties in cleaving the DsbC fusion tag. This may be true for peptides containing a proline, isoleucine or valine at the N-terminus, as data presented in the previous chapter revealed that

TEV protease was not completely efficient in cleaving sequences containing those residues at position P1´. For this subset of peptides inclusion of an extra glycine residue at the N-terminus might lead to a higher recovery of cleaved peptide. However, it is also possible that presence of this extra residue will affect the biological activity of the resulting recombinant animal venom peptides.

6.5. Conclusion

The natural venom collection may be comprised of up to 40,000,000 different peptides and can be seen as an attractive source of stable and evolutionarily fine-tuned highly selective molecules for drug discovery. In the scope of the VENOMICS project a sequence database of ~25,000 novel genes encoding venom peptides was produced based on transcriptomic and proteomic data collected in 201 different animal species. A subset of ~5000 peptides with more than 35 amino acids and representative of the natural diversity observed in the database was selected for recombinant production in bacteria following protocols established in the previous chapter. Here we describe the high-throughput production of a unique library of recombinant venom peptides expressed in *E. coli*. The library contains 2736 animal venom peptides that were shown to be properly oxidized. Recently, the library described here was screened on several therapeutic targets. The data allowed identification of dozens of different hits in the majority of the molecular targets tested. Overall, data presented here confirms that *E. coli* is an effective host for the production of large libraries of venom peptides of remarkable interest in drug discovery programs. While the exact sequence of the 4992 peptides of this study remains confidential, the high-throughput protein production process described here could be adapted to the generation of innovative libraries of different peptides and protein families with interest for drug discovery.

7. GENERAL DISCUSSION AND FUTURE PERSPECTIVES

Animal venoms are complex chemical cocktails comprising a wide range of biologically active reticulated peptides that target, with high selectivity and efficacy, a variety of membrane receptors (Lewis & Garcia, 2003). Venoms can, therefore, be seen as large natural libraries of functional molecules that are continuously being selected and highly refined by the evolution process, up to the point where every molecule is endowed with pharmacological properties that are highly valuable in the context of human use and drug development. Recently, considerable emphasis is being put on the discovery and characterization of venom peptides, using animal venoms as a source of potential drugs. Reflecting the complexity of venoms, the number of venom peptides that remain to be discovered may be remarkable large. Considering the number of venomous animal species, the global animal venom resource can be seen as a collection of millions of different molecules of which only a small part is known (Escoubas & King, 2009; King, 2011). Thus, use of venoms for drug discovery is rapidly emerging but remains mostly an unrealized prospective, due to several major difficulties related with venom research, including the availability of material, sample size (most venomous animals are small to very small), the complex nature of their composition and capacity to produce them recombinantly (Vetter *et al*., 2011).

Traditionally, research on new toxins to identify novel therapeutics has been based in the individual characterization of a venom peptide, making the identification of a drug lead from one or few molecules a needle-in-a-haystack problem. The classical bioassay-guided isolation of bioactive peptides approach is time consuming, risky and not applicable to tens to hundreds of venoms in parallel. To try solving this issue and explore more rationally the pharmacological potential of venom peptides, the European consortium VENOMICS was set up, offering a totally new paradigm that completely bypasses the classical approach to identify novel therapeutics within venoms. To explore in depth the diversity of these animal toxins, a platform for the efficient production of a large panel of natural venom peptides was developed. This lead to the improvement of high-throughput protocols that comprise transcriptomics and proteomics analysis, chemical and recombinant production of venom peptides, and high-fidelity assays for functional analysis of drug leads. This doctoral research work was integrated in the activities of VENOMICS project, particularly in the development of methodologies for the efficient production of functional venom peptides recombinantly in *E. coli*. Specifically, the main goal of the work described here was to develop a high-throughput gene synthesis platform to produce synthetic genes for the production of large recombinant venom peptide libraries and to develop accessory research on the parameters related with gene design and synthesis. This project culminated with the production of ~5,000 synthetic genes that allowed building the largest ever generated library of recombinant venom peptides for drug discovery.

Development of innovative, high-throughput, molecular biology methods to produce synthetic genes is crucial for the rapid advancement of the drug discovery pipeline. The third chapter of this thesis fine-tuned a novel procedure to obtain synthetic genes and described the development of a robust and efficient high-throughput gene synthesis platform. Large transcriptomic and proteomic studies established within VENOMICS lead to the identification of about 25,000 novel venom peptides from 201 different animal species. Although novel peptide sequences become known at a tremendous pace, the genetic material required for their production is difficult to obtain. Gene synthesis allows producing nucleic acids when pre-existing material does not exist. In addition, it allows applying gene design strategies to optimize DNA sequences to the heterologous host. Chapter 3 reported the strategies developed to design and optimize the gene sequences of small genes (<0.5 kb) for high levels of expression in *E. coli*, and to synthesise multiple high-fidelity gene targets encoding venom peptides simultaneously. A novel algorithm for codon optimization was developed to improve recombinant expression of venom peptides in *E. coli*. The algorithm developed here (ATGenium) allowed designing optimized genes that mimic the natural gene properties which are relevant for recombinant expression. Codon usage is an important parameter (as described below) for gene design and usually takes in consideration the codon frequency for each amino acid in highly expressed genes of the selected host system. One source of general debate is whether the codons for a given amino acid should be selected randomly or simply should correspond to codons with higher frequency for the host system selected. This latter approach corresponds to the traditional strategy that aims maximizing CAI values, which is still used in some codon optimization algorithms. In our study, we used an initial strategy to ensure that there is no imbalance between the number of tRNA molecules within bacterial cells and their requirement by the translational machinery. This aims avoiding the overuse of the more frequent codons for a given amino acid. Here, gene design of DNA sequences was performed with high efficiency using the ATGenium algorithm that not only randomly selects codons from an optimized codon usage table for *E. coli*, but also maximizes the formation of stable mRNA molecules. In addition, ATGenium minimizes the presence of repeated sequences, avoids the appearance of *E. coli* regulatory sequences and fixes GC content to optimal values. In contrast to the majority of commercial solutions currently available, this algorithm was integrated within a bioinformatics tool to work in a high-throughput fashion, allowing simultaneously designing multiple genes in a simple and very rapid manner.

To support the construction of a robust and efficient gene synthesis method, different approaches were investigated aiming ensuring simplicity and speed in synthetic gene production. A PCR-assembly methodology (PCA-DTF method), using *Kod* DNA polymerase under optimized conditions, was selected as the most effective strategy to produce synthetic genes. This novel approach improves the efficiency of the PCR-assembly reaction and

reduces the cost associated with the acquisition of oligonucleotides, since the use of primers with 60 nt in length, with 20 nt overlaps and a gap of 20 nt, decreases the number of primers required to obtain a given gene. In addition, we observed the highest percentage of clones without errors in genes synthesised with oligonucleotides without purification (desalted). In contrast to the expectation, oligonucleotide purifications did not improve the fidelity of gene products meaning that the purification methods analysed did not avoid the appearance of mutations within synthetic oligonucleotides. Gene sequence analysis revealed that the most frequent mutations identified in synthetic genes obtained in Chapter 3 are single deletion, which represent 44% of all mutations similarly to what was described in other studies (Fuhrmann *et al*., 2005; Saaem *et al*., 2012; G. Wu *et al*., 2006). The direct cloning of synthetic gene products into the expression vector was also crucial for the implementation of the novel platform. Usually, synthetic genes are cloned into a standard cloning vector and then transferred to the expression vector; direct integration into the expression vector leads to a higher efficacy of the process. As a proof of concept that this platform would be effective to synthesise multiple genes, simultaneously and with high-fidelity, 96 venom peptide sequences were used to design 96 optimized genes for high expression levels in *E. coli*. The data revealed that this HTP platform presents high efficiencies for gene assembly through PCR and efficient cloning into *E. coli* expression vector. The error rate of the large scale synthetic production was 1.1 errors per kb, showing that for the identification of an error-free gene product it is necessary only to screen a maximum of three clones. This error rate is similar to those obtained by other methods (Xinxin Gao *et al*., 2003; Hoover & Lubkowski, 2002; G. Wu *et al*., 2006; Xiong *et al*., 2004; Zampini *et al*., 2015). However, these methods are described for single gene production protocols and there is no data reporting multiple and simultaneously gene production in the literature.

The main limitation for high fidelity gene synthesis remains the quality of synthetic oligonucleotides. To ensure 100% gene synthesis accuracy more efficient methods to produce synthetic oligonucleotides are required. However, for large scale production there must be a commitment between simplicity, speed and cost. Thus, we believe that the gene synthesis approach described here is an accurate, low cost and rapid system to produce synthetic genes encoding venom peptides, although the fidelity of gene products does not equal 100%. The results obtained for the high-throughput synthesis of 96 genes simultaneously emphasize the success of this approach, since the same conditions were used for all genes, which often have different features, such as GC content, that can affect the efficiency of gene synthesis protocol. In addition, the novel HTP gene synthesis platform could be adapted to synthesise DNA molecules from different origins to be expressed in any host system, allowing recombinant production of functional proteins to be used in a variety of applications. Although an efficient system has been developed in this work for large scale production of error-free synthetic genes

with small lengths, it is well known that there are difficulties related with the synthesis of large genes since in these cases there is an increased probability of incorporating errors during PCR-assembly. High error rates in artificial gene synthesis are regarded as the biggest obstacle to produce high-fidelity genes. A recent study compared different gene synthesis methods and showed that the error rates can vary from 1 error per 100 bases to 1 error to 1000 bases. These values could be reduced to 1 error per 10,000 bases when error-correction techniques are used (Kosuri & Church, 2014). Several authors have explored different methods to reduce error rates, including strategies based on additional steps for error removal from synthetic genes. The use of enzymes that recognize DNA mismatches is suggested to be an interesting strategy for removing errors that are copied and accumulated during gene synthesis procedures (Carr, 2004; Fuhrmann *et al*., 2005; Saaem *et al*., 2012; Wan *et al*., 2014).

The production of longer synthetic genes (with 1 or more kb in length) requires a higher number of oligonucleotides as starting material, which dramatically increases the probability of appearance of errors within the final sequence. In Chapter 4, we described an integrated gene synthesis method that includes an additional step to remove mutations accumulated in synthetic genes, using mismatch cleavage enzyme T7 endonuclease I. Mismatch cleavage enzymes have the ability to recognize incorrect impairments between DNA strands and cleave DNA fragments near to the DNA mismatch. Thus, this group of enzymes is adequate to decrease the occurrence of mutations in synthetic gene products. The performance of six endonuclease enzymes, from different sources, to specifically recognize and cleave DNA mismatches was analysed here. Initially, the proteins were expressed in *E. coli* cells and purified by IMAC. The data revealed that only bacterial proteins were expressed in soluble form, while eukaryotic proteins were accumulated in the insoluble fraction. Out of the recombinant endonucleases produced, only T7 endonuclease I derivatives showed nuclease activity. Thus, the efficacy of recombinant T7 endonuclease I derivatives (6HIS and MBP) to recognize and cleave DNA fragments with mismatches was evaluated during the synthesis of an artificial *gfp* gene. This gene was synthesised with success using an integrated gene synthesis method, which includes two consecutives PCR reactions, followed by an enzymatic mismatch cleavage step (EMC) with T7 endonuclease I for error removal, and a third PCR reaction to recover the high fidelity gene. Data from functional and sequence analysis of synthetic *gfp* gene products revealed that the number of mutations was strongly reduced when T7 endonuclease I was used. Moreover, the results obtained showed that the mutation frequency was dramatically reduced by 8-fold relative to gene synthesis not incorporating an error repair step, resulting in an error frequency of 0.43 errors per kb. Untreated genes presented a higher frequency of deletions while this mutation was reduced in synthetic genes

exposed to the enzyme treatments. For instance, error repair with T7 endonuclease I-MBP reduced the presence of single deletions by 17-fold.

Proofreading activity during gene synthesis is related with a combined action of T7 endonuclease I and *Kod* Hot Start DNA polymerase that has a strong 3'-5' exonuclease action, making the integrated gene synthesis method described here particularly robust and suitable to reduce mutations in artificial genes. This approach reduces the dependence of gene synthesis fidelity on the quality of oligonucleotides used as initial templates for PCR assembly, since mutations can be removed after gene synthesis using an enzymatic treatment with T7 endonuclease I. Although the inclusion of an error correction step with T7 endonuclease I was effective in reducing mutations observed in synthetic genes, data presented here confirm that this error removal step is unable to completely abolish the number of errors in final DNA sequences. This result can be explained by inefficient hetero-duplexes formation before the error correction step. In addition, the errors identified in artificial genes can be introduced during the third PCR, which uses the outer primers that can include errors from imperfect chemical synthesis. Globally, the work reveals the capacity of T7 endonuclease I to improve the fidelity of gene synthesis, opening avenues to explore the high potency of gene synthesis technologies. The benefits of using T7 endonuclease I relate with the capacity of easily producing this enzyme in *E. coli*, but also with the simplicity and speed of the enzymatic treatment step that can be easily integrated in high-throughput platforms. Future work should be performed in order to explore the specific activity of T7 endonuclease I, particularly to improve its capacity to recognize deletions, insertions or substitutions. The activity of T7 endonuclease I should be analysed for correction of synthetic genes with specific and known errors.

In chapter 5, the influence of different factors, such as gene design, presence of fusion tags and usage of TEV protease for tag removal, on the recombinant expression of disulfide-rich venom peptides in *E. coli* was investigated. The importance of codon usage for the expression of artificial genes encoding animal venom peptides in *E. coli* was analysed. *De novo* gene synthesis is the most convenient way to obtain genes for recombinant expression. This is particularly true for genes encoding venom peptides for which native biological material is usually not available as palpable DNA. Designing a gene to express the target protein requires choosing from an enormous number of possible DNA sequences (Welch *et al*., 2011). In this thesis we have investigated the effects of codon usage on levels of 24 purified venom peptides, with various lengths and containing different number of disulfide bridges, obtained from *E. coli*. Levels of purified proteins are intimately related with gene expression and the data suggest that gene design is intimately related with transcription efficacy. However, correlations between primary sequence of gene variants and properties that have been suggested to affect

expression, such as the number of disulfide bridges, peptide size, CAI value or GC content (Welch, Villalobos, *et al*., 2009) were not significant, suggesting that these factors do not directly affect gene expression. Thus, it was suggested that the differences observed in gene expression might be explained by codon usage, in particular by subtle differences in DNA sequences generated by random selection of codons for a given amino acid. The data showed that high expresser gene variants produced up to twice the levels of recombinant protein when compared with low expressing ones. Comparisons between high and low expressers suggested that factors such as the frequency of cysteine codons could explain differences between expression levels. Thus, the data presented here reveal that high levels of expression of venom peptides require a similar usage of the two cysteine codons, Cys-TGT and Cys-TGC. It is now well established that high translation rates contribute to deplete the cellular translational machinery (Dong *et al*., 1995). Considering the particular case of venom peptides, cysteine is a highly frequent residue being at least four times more frequent in the recombinant fusion genes than in *E. coli*. Thus, recombinant *E. coli* strains expressing venom peptides at high levels require a similar usage of both cysteine codons most possibly to avoid depletion of one relative to the other. The data suggest that if one codon is present in higher frequencies, then it will be more easily depleted within the cell and will become the limiting codon for rate of gene synthesis. Thus, codon usage seems to play an important role for high expression yield of recombinant disulfide-rich peptides. Systematic analysis of the relationship between gene sequences and expression levels will be a powerful tool to refine future design algorithms.

Chapter 5 described the generation of a set of five novel vectors to improve the levels of venom peptides expressed in the soluble form in *E. coli*. The vectors allow the fusion of the encoded recombinant peptide to a protein partner (tag) that is known to be expressed at high levels and in the soluble form in *E. coli*. Although *E. coli* is a highly robust bioreactor for heterologous protein expression, the production of disulfide-bonded proteins in these bacteria is hampered by the lack of an effective post-translational system, which is often required to effectively express recombinant eukaryotic proteins. Here, we describe the influence of two different fusion partners, disulfide-bond isomerase (DsbC) and maltose binding protein (MBP), which display redox and no redox properties, respectively, in the recombinant expression of functional venom peptides in *E. coli*. The novel pHTP vectors created here insert the engineered fusion tags at the N-terminus of the recombinant proteins. The cloned genes are controlled by regulated T7 promoters and sequences encoding tags selected for incorporating in the vectors were previously shown to be highly effective in raising levels of protein expression and solubility. Two of the vectors which encode fusion partners contain a signal peptide to target venom peptide expression into the periplasmic compartment (pHTP4, pHTP6). The remaining fusion tags will lead to cytoplasmic recombinant protein expression. After development of the different pHTP vector derivatives, a comparison study was

implemented to compare the capacity of 5 different fusion partners to drive the expression of 16 recombinant venom peptides, with different origins and cysteine bond patterns, in BL21(DE3) pLysS *E. coli* cells. The data showed that the recombinant peptides displayed different degrees of expression and solubility. Depending on the peptide and the fusion used, the levels of purified fusion peptides varied from zero (pHTP1 vector) to more than 300 mg of purified protein per liter of culture. Data presented here revealed that the best way to express high yields of folded animal venom peptides is their expression in the periplasmic compartment of *E. coli* cells. Thus, presence of the signal peptide leads to higher levels of expression for both the DsbC (pHTP4) and the MBP (pHTP6) fusion tags. The positive results obtained with DsbC may be explained by its excellent solubilization potential but also by its isomerase and chaperon activities, which promotes the correct folding of venom peptides. Data from LC-MS also confirmed the correct oxidation state of the final recombinant peptide indicating that the fusion peptides produced with pHTP4 (DsbC) exported to the periplasmic cellular compartment folded correctly.

Removal of fusion tags is fundamental to fully rescue functional recombinant animal venom peptides. In the specific case of venom peptides, it is known that the N-terminal part of the peptide comprises residues involved in receptor binding (Karbat *et al*., 2007). Thus, presence of an N-terminal fusion tag at the N-terminus of venom peptides may affect their biological activity. In this study, the protease selected to remove fusion tags was the TEV protease. Cleavage activity of this protease requires a glycine (Gly) or a serine (Ser) residue at the C-terminus of its recognition site (Dougherty *et al*., 1988), leaving non-native Ser and Gly at the N-terminus of the target peptide after tag removal. This may critically affect binding efficacy of recombinant peptide. Thus, the performance of TEV protease was analysed in non-optimal conditions for TEV proteolysis (buffer conditions favorable to produce recombinant venom peptides) and when the P1' position of the protease recognition site was other than Ser or Gly. Data collected here suggest that, with the exception of proline, all other residues can be accommodated at P1' position of the TEV protease recognition site without notably affecting the efficacy of TEV proteolysis, as described by Kapust *et al*. (2002). Thus, the cleavage activity of TEV protease allows recovering venom peptides with the same sequence properties than native molecules, allowing the biological function of peptides to be retained.

The experiments of chapter 6 explored the utilization of innovative and robust methodologies, which were optimized in the previous chapters (chapter 3, 4 and 5), to explore the huge biodiversity presented by animal venoms. The application of the findings reported above allowed building a high-throughput platform for the efficient production of thousands of animals' venom peptides in *E. coli*. Although the use of venoms for drug discovery is rapidly emerging, the full exploitation of venom peptide potential is still undefined due to several technical

bottlenecks, including the capacity to recombinantly produce these highly relevant molecules in an efficient and simple manner. The results presented in chapters 3 and 5 showed that it is possible to overcome the technical limitations related firstly with the difficulty in obtaining biological material in sufficient amounts, since the efficient and high-fidelity production of synthetic genes allowed to generate any DNA sequence optimized for expression in *E. coli*. On the other hand, to overcome the disadvantages related with the production of disulfide rich peptides in the cytoplasm of *E. coli* cells, a system was developed for the efficient production of disulfide rich peptides in this bacterium. This platform included the appropriate design of genes encoding venom peptides and an expression vector that carry the most efficient fusion partner (DsbC), not only to improve the expression levels but also to ensure that the recombinant peptides assumed proper folding.

As a challenging task of VENOMICS project, in chapter 6 we described the use of the automated high-throughput gene synthesis platform developed here to produce 4992 synthetic genes encoding venom peptides. The primary sequence of 4992 reticulated peptides, originated from 201 venomous animal species, were used to design genes using the ATGenium algorithm described in chapter 3, using an optimized codon usage that includes the same usage of the two cysteine codons, Cys-TGT and Cys-TGC. Genes with an average size of 220 bp, an average GC content of 49% and an average CAI of 0.92 were sliced into oligonucleotides with an overlap region of 20 bp and a gap of 20 bp, having a maximum length of 60 bp. The optimized genes were designed for high expression in *E. coli* and were synthesised using a PCR-based method to assemble overlapping oligonucleotides using the optimal conditions identified in chapter 3. Individual genes were cloned into pHTP4 expression vector, which contains DsbC fusion partner. The presence of a signal peptide allows the exportation of recombinant peptides to the periplasm, thus favoring the formation of disulfide bonds that are required to obtain fully functional and stable venom peptides. The robustness of the gene synthesis pipeline was demonstrated by the integrity of resulting artificial nucleic acids. The data revealed that the majority of genes (3818, 76.5%) are correct when only one clone is screened. Out of 1174 genes for which no correct clones were obtained in the first screening, 809 of the genes (16.2%) were found to be correct when a second clone was screened. For the remaining 365 genes it was necessary to screen a third clone to identify a correct acid nucleic. Thus, these results revealed that it was necessary to screen an average of 1.3 clones to obtain a synthetic gene with the correct sequence. Taken together, the data showed that the error rate of this gene synthesis platform is 1.06 errors/kb, which is a low error rate and it is similar or slightly better when compared with other gene synthesis methodologies based on PCR or ligation assembly (Ma *et al*., 2012; Saaem *et al*., 2012; G. Wu *et al*., 2006). Clearly, it is difficult to compare error rates obtained by other methods with the data obtained in the HTP gene synthesis platform described here, since the error rates described in the

literature are related with synthesis of single genes. Instead, the HTP gene synthesis platform described here was used to simultaneously produce pools of 96 genes, using the same conditions for all genes, which are different in length, GC content and number of cysteines residues. As observed previously, incorrect bases are likely to be incorporated into the artificial genes during PCR assembly. The majority of the errors identified during the production of 4992 genes were deletions (76%), as it is expected from the chemical synthesis of oligonucleotides (LeProust *et al*., 2010). This result is in accordance with other studies (Carr, 2004; Ma *et al*., 2012; Saaem *et al*., 2012) and it is similar to results obtained and described in chapter 3. Thus, the most common errors observed during *de novo* gene synthesis are deletions that result from the inclusion of truncated versions of the oligonucleotides used as a starting material in PCR assembly reaction. In general, these errors require an error removal step to retrieve the correct sequence (Ma *et al*., 2012; Xiong *et al*., 2008). However, this process introduces complexity in the gene synthesis process, which most of times is not compatible with high-throughput and automated methods. Since genes encoding venom peptides are relatively small (average size of 220 bp) and taking in account the results presented in chapter 3, a low error rate was expected and, therefore, an error correction technology was not employed in this pipeline. In summary, a total of 1100 kb of DNA were synthesised with success during this work. The entire gene synthesis process to produce 4×96 synthetic genes encoding venom peptides can be completed in 6-7 days. This procedure represents a low cost and accurate high-throughput method with a low error rate to generate artificial nucleic acids. Progress in large-scale and low-cost construction of desired DNA sequences confer a number of unique advantages for both fundamental and applied biological research. The ability to *de novo* produce any DNA sequence increases the number of biological hypothesis that can be analysed reducing the labor time required for the construction of a desired DNA sequence.

In conclusion, this thesis describes a totally novel approach to produce small synthetic genes encoding venom peptides. A HTP platform was developed to produce pools of dozens to thousands genes optimized to obtain high expression levels in *E. coli*, using simple and fast methods to support the large scale production. In addition, different parameters that are known to affect the recombinant expression were investigated, defining appropriated methods to design genes and for removal of fusion partners, which are required to ensure high levels of soluble peptides with proper folding. Finally, all findings obtained during this research project were used for the large-scale production of 4992 synthetic genes encoding venom peptides. The work described in this thesis generated the biggest library of synthetic genes encoding venom peptides constructed to date. Data from recombinant expression in *E. coli* obtained by a collaborator laboratory revealed that > 50% of venom peptides were produced in soluble form, which also emphasizes the success of the novel HTP gene synthesis platform. Thus, the limitations related with the recombinant production of venom peptides in *E. coli* were, in part,

overcome. The venom peptide bank produced within VENOMICS is a powerful tool for drug discovery. This library can be directly used by pharmaceutical industry to identify new molecules with therapeutic interest. Future work should address the possibility of adjusting different codon usage strategies to the primary sequence of the target proteins. In addition, studies with the pHTP expression vectors should be extended, in particular to analyse the effect of novel tags in levels of recombinant peptide expression and their efficacy to improve the formation of disulfide bonds.

8. BIBLIOGRAPHIC REFERENCES

Abdullah, J., Joachimiak, A., & Collart, F. (2009). “System 48” High-throughput cloning and protein expression analysis. *Methods in Molecular Biology*, *498*, 117–127. http://doi.org/10.1007/978-1-59745-196-3.

Agarwal, K. L., Buchi, H., Caruthers, M. H., Gupta, N., Khorana, H, G., Kleppe, K., & Kumar, A. (1970). Total synthesis of the gene for an alanine transfer ribonucleic acid from yeast. *Nature*, *227*, 27–34.

Aili, S. R., Touchard, A., Escoubas, P., Padula, M. P., Orivel, J., Dejean, A., & Nicholson, G. M. (2014). Diversity of peptide toxins from stinging ant venoms. *Toxicon*, *92*, 166–178. http://doi.org/10.1016/j.toxicon.2014.10.021.

Allert, M., Cox, J. C., & Hellinga, H. W. (2010). Multifactorial determinants of protein expression in prokaryotic open reading frames. *J. Mol. Biol.*, *402*, 905–918.

Altschul, S. F., Madden, T. L., Schäffer, a a, Zhang, J., Zhang, Z., Miller, W., & Lipman, D. J. (1997). Gapped BLAST and PSI-BLAST: a new generation of protein database search programs. *Nucleic Acids Research*, *25*(17), 3389–402.

Anangi, R., Rash, L. D., Mobli, M., & King, G. F. (2012). Functional Expression in *Escherichia coli* of the Disulfide-Rich Sea Anemone Peptide APETx2, a Potent Blocker of Acid-Sensing Ion Channel 3. *Marine Drugs*, *10*(7), 1605–1618. http://doi.org/10.3390/md10071605.

Aravind, L., Makarova, K. S., & Koonin, E. V. (2000). Holliday junction resolvases and related nucleases: identification of new families, phyletic distribution and evolutionary trajectories. *Nucleic Acids Research*, *28*(18), 3417–3432. http://doi.org/10.1093/nar/28.18.3417.

Ashman, K., Matthews, N., & Frank, R. W. (1989). Chemical synthesis, expression and product assessment of a gene coding for biologically active human tumour necrosis factor alpha. *Protein Engineering*, *2*(5), 387–91.

Babon, J. J., McKenzie, M., & Cotton, R. G. H. (2003). The use of resolvases T4 endonuclease VII and T7 endonuclease I in mutation detection. *Molecular Biotechnology*, *23*(1), 73–81. http://doi.org/10.1385/MB:23:1:73.

Baker, R. T. (1996). Protein expression using ubiquitin fusion and cleavage. *Current Opinio in Biotechnology*, *7*, 541–546.

Balbás, P. (2001). Understanding the art of producing protein and nonprotein molecules in *Escherichia coli*. *Molecular Biotechnology*, *19*(3), 251–67. http://doi.org/10.1385/MB:19:3:251.

Baneyx, F. (1999). Recombinant protein expression in *Escherichia coli*. *Current Opinion in Biotechnology*, *10*, 411–421.

Barany, F., & Gelfand, D. H. (1991). Cloning, oeverexpression and nucleotide sequence of a thermostable DNA ligase-encoding gene. *Gene*, *109*, 1–11.

Bazaa, A., Marrakchi, N., El Ayeb, M., Sanz, L., & Calvete, J. (2005). Snake venomics: comparative analysis of the venom proteomes of the Tunisian snakes Cerastes cerastes, Cerastes vipera and Macrovipera libetina. *Proteomics 5*, 4223–4235.

Bende, N. S., Dziemborowicz, S., Herzig, V., Ramanujam, V., Brown, G. W., Bosmans, F., … Mobli, M. (2015). The insecticidal spider toxin SFI1 is a knottin peptide that blocks the pore of insect voltage-gated sodium channels via a large b-hairpin loop. *FEBS Journal*, *282*(5), 904–920. http://doi.org/10.1111/febs.13189.

Bende, N. S., Dziemborowicz, S., Mobli, M., Herzig, V., Gilchrist, J., Wagner, J., … Bosmans, F. (2014). A distinct sodium channel voltage-sensor locus determines insect selectivity of the spider toxin Dc1a. *Nature Communications*, *5*, 4350. http://doi.org/10.1038/ncomms5350.

Benner, S., & Sismour, A. (2005). Synthetic biology. *Nature Reviews. Genetics.*, *6*(July), 533–543. http://doi.org/10.1038/nrg1637.

Berkmen, M. (2012). Production of disulfide-bonded proteins in *Escherichia coli*. *Protein Expression and Purification*. http://doi.org/10.1016/j.pep.2011.10.009.

Berkmen, M., Riggs, P., Faulkner, M., Jeans, C., Emrich, C. A., & Lobstein, J. (2012). SHuffle, a novel *Escherichia coli* protein expression strain capable of correctly folding disulfide bonded proteins in cytoplasm. *Microbial Cell Factories*, *11*(56).

Berrow, N. S., Alderton, D., & Owens, R. J. (2009). The precise engineering of expression vectors using high-throughput In- Fusion PCR cloning. *Methods in Molecular Biology*, *498*, 75–90.

Bershtein, S., & Tawfik, D. S. (2008). Advances in laboratory evolution of enzymes. *Current Opinion in Chemical Biology*, *12*, 151–158. http://doi.org/10.1016/j.cbpa.2008.01.027.

Bessette, P. H., Aslund, F., Beckwith, J., & Georgiou, G. (1999). Efficient folding of proteins with multiple disulfide bonds in the *Escherichia coli* cytoplasm. *Proceedings of the National Academy of Sciences of the United States of America*, *96*(24), 13703–8.

Binkowski, B. F. (2005). Correcting errors in synthetic DNA through consensus shuffling. *Nucleic Acids Research*, *33*(6), e55–e55. http://doi.org/10.1093/nar/gni053.

Binkowski, B. F., Richmond, K. E., Kaysen, J., Sussman, M. R., & Belshaw, P. J. (2005). Correcting errors in synthetic DNA through consensus shuffling. *Nucleic Acids Research*, *33*(6), e55. http://doi.org/10.1093/nar/gni053.

Blommel, P. G., & Fox, B. G. (2007). A combined approach to improving large-scale production of tobacco etch virus protease. *Protein Expression and Purification*, *55*(1), 53–68. http://doi.org/10.1016/j.pep.2007.04.013.

Braun, P., Hu, Y., Shen, B., Halleck, A., Koundinya, M., Harlow, E., & LaBaer, J. (2002). Proteome-scale purification of human proteins from bacteria. *Proceedings of the National Academy of Sciences of the United States of America*, *99*(5), 2654–9. http://doi.org/10.1073/pnas.042684199.

Bray, G. (2006). Exenatide. *Am. J. Health. Syst. Pharm.*, *63*(5), 411–418.

Bruni, R., & Kloss, B. (2013). High-throughput cloning and expression of integral membrane proteins in *Escherichia coli*. *Current Protocols in Protein Science*, *74*, 29.6.1–29.6.34. http://doi.org/10.1002/0471140864.ps2906s74.

Bryksin, A. V, & Matsumura, I. (2010). Overlap extension PCR cloning: a simple and reliable way to create recombinant plasmids. *BioTechniques*, *48*(6), 463–5. http://doi.org/10.2144/000113418.

Buczek, O., Bulaj, G., & Olivera, B. M. (2005). Conotoxins and the posttranslational modification of secreted gene products. *Cellular and Molecular Life Sciences*, *62*(24), 3067–3079. http://doi.org/10.1007/s00018-005-5283-0.

Butt, T. R., Edavettal, S. C., Hall, J. P., & Mattern, M. R. (2005). SUMO fusion technology for difficult-to-express proteins. *Protein Expression and Purification*, *43*(1), 1–9. http://doi.org/10.1016/j.pep.2005.03.016.

Calvete, J. J., Sanz, L., Angulo, Y., Lomonte, B., & Gutiérrez, J. M. (2009). Venoms, venomics, antivenomics. *FEBS Letters*, *583*(11), 1736–1743. http://doi.org/10.1016/j.febslet.2009.03.029.

Cao, Z., Yu, Y., Wu, Y., Hao, P., Di, Z., He, Y., … Li, W. (2013). The genome of Mesobuthus martensii reveals a unique adaptation model of arthropods. *NATURE COMMUNICATIONS*, *4*.

Cardoso, F. C., Dekan, Z., Rosengren, K. J., Erickson, A., Vetter, I., Deuis, J. R., … Lewis, R. J. (2015). Identification and Characterization of ProTx-III [μ-TRTX-Tp1a], a New Voltage-Gated Sodium Channel Inhibitor from Venom of the Tarantula Thrixopelma pruriens. *Molecular Pharmacology*, *88*(2), 291–303. http://doi.org/10.1124/mol.115.098178.

Carr, P. A. (2004). Protein-mediated error correction for de novo DNA synthesis. *Nucleic Acids Research*, *32*(20), e162–e162. http://doi.org/10.1093/nar/gnh160.

Carr, P. A., & Church, G. (2009). Genome engineering. *Nature Biotechnology*, *27*(12), 1151–1162. http://doi.org/10.1038/nbt1590.

Caruthers, M. H., Barone, A. D., Beaucage, S. L., Dodds, D. R., Fisher, E. F., McBride, L. J., … Tang, J. Y. (1987). Chemical synthesis of deoxyoligonucleotides by the phosphoramidite method. *Methods in Enzymology*, *154*, 287–313.

Caruthers, M. H., Beaucage, S. L., Becker, C., Efcavitch, J. W., Fisher, E. F., Galluppi, G., … McBride, L. (1983). Deoxyoligonucleotide synthesis via the phosphoramidite method. *Gene Amplification and Analysis*, *3*, 1–26.

Casewell, N. R., Wüster, W., Vonk, F. J., Harrison, R. A., & Fry, B. G. (2013). Complex cocktails: the evolutionary novelty of venoms. *Trends in Ecology & Evolution*, *28*(4), 219–229. http://doi.org/10.1016/j.tree.2012.10.020.

Catterall, W. A. (1995). Structure and Function of Voltage-Gated Ion Channels, 493–531.

Chambers, S. P. (2002). High-throughput protein expression for the post-genomic era. *Drug Discovery Today*, *7*(14), 759–65.

Chang, Z., Lu, M., Ma, Y., Kwag, D.-G., Kim, S.-H., Park, J.-M., … Park, J.-S. (2015). Production of disulfide bond-rich peptides by fusion expression using small transmembrane proteins of *Escherichia coli*. *Amino Acids*, *47*(3), 579–87. http://doi.org/10.1007/s00726-014-1892-y.

Chao, R., Yuan, Y., & Zhao, H. (2014). Recent advances in DNA assembly technologies. *FEMS Yeast Research*, 1–11. http://doi.org/10.1111/1567-1364.12171.

Cheng, J.-Y., Chen, H.-H., Kao, Y.-S., Kao, W.-C., & Peck, K. (2002). High throughput parallel synthesis of oligonucleotides with 1536 channel synthesizer. *Nucleic Acids Research*, *30*(18), e93.

Cheung, R. C. F., Wong, J. H., & Ng, T. B. (2012). Immobilized metal ion affinity chromatography: A review on its applications. *Applied Microbiology and Biotechnology*, *96*(6), 1411–1420. http://doi.org/10.1007/s00253-012-4507-0.

Choi, J. H., & Lee, S. Y. (2004). Secretory and extracellular production of recombinant proteins using *Escherichia coli*. *Applied Microbiology Biotechnology*, *64*(5), 625–635. http://doi.org/10.1007/s00253-004-1559-9.

Chong, S., Mersha, F. B., Comb, D. G., Scott, M. E., Landry, D., Vence, L. M., … Xu, M. Q. (1997). Single-column purification of free recombinant proteins using a self-cleavable affinity tag derived from a protein splicing element. *Gene*, *192*(2), 271–81.

Clement, H., Flores, V., Diego-Garcia, E., Corrales-Garcia, L., Villegas, E., & Corzo, G. (2015). A comparison between the recombinant expression and chemical synthesis of a short cysteine-rich insecticidal spider peptide. *The Journal of Venomous Animals and Toxins Including Tropical Diseases*, *21*, 19. http://doi.org/10.1186/s40409-015-0018-7.

Collins-Racie, L. A., McColgan, J. M., Grant, K. L., DiBlasio-Smith, E. A., McCoy, J. M., LaVallie, E. R. (1995). Production of recombinant bovine enterokinase catalytic subunit in *Escherichia coli* using the novel secretory fusion partner DsbA. *Biotechnology (N. Y.)*, *13*(9), 982–987.

Colwill, K., Wells, C. D., Elder, K., Goudreault, M., Hersi, K., Kulkarni, S., … Morin, G. B. (2006). Modification of the Creator recombination system for proteomics applications--improved expression by addition of splice sites. *BMC Biotechnology*, *6*, 13. http://doi.org/10.1186/1472-6750-6-13.

Costa, S. J., Almeida, A., Castro, A., & Domingues, L. (2014). Fusion tags for protein solubility, purification and immunogenicity in *Escherichia coli*: the novel Fh8 system. *Frontiers in Microbiology*, *5*(February), 63. http://doi.org/10.3389/fmicb.2014.00063.

Costa, S. J., Almeida, A., Castro, A., Domingues, L., & Besir, H. (2013). The novel Fh8 and H

fusion partners for soluble protein expression in *Escherichia coli*: a comparison with the traditional gene fusion technology. *Applied Microbiology and Biotechnology*, *97*(15), 6779–91. http://doi.org/10.1007/s00253-012-4559-1.

Craik, D. J., Fairlie, D. P., Liras, S., & Price, D. (2013). The Future of Peptide-based Drugs. *Chemical Biology & Drug Design*, *81*(1), 136–147. http://doi.org/10.1111/cbdd.12055.

Currin, A., Swainston, N., Day, P. J., & Kell, D. B. (2014). SpeedyGenes: an improved gene synthesis method for the efficient production of error-corrected, synthetic protein libraries for directed evolution. *Protein Engineering Design and Selection*, *27*(9), 273–280. http://doi.org/10.1093/protein/gzu029.

Cushman, D., & Ondetti, M. (1991). History of the design of captopril and related inhibitors of angiotensin converting enzyme. *Hypertension*, *17*, 589–592.

Czar, M. J., Anderson, J. C., Bader, J. S., & Peccoud, J. (2009). Gene synthesis demystified. *Trends in Biotechnology*, *27*(2), 63–72. http://doi.org/10.1016/j.tibtech.2008.10.007.

Davis, G. D., Elisee, C., Newham, D. M., Harrison, R. G. (1999). New fusion protein systems designed to give soluble expression in *Escherichia coli*. *Biotechnology and Bioengineering*, *8*, 1668–1674.

Davis, J., Jones, A., & Lewis, R. J. (2009). Remarkable inter- and intra-species complexity of conotoxins revealed by LC/MS. *Peptides*, *30*(7), 1222–1227. http://doi.org/10.1016/j.peptides.2009.03.019.

de Marco, A. (2009). Strategies for successful recombinant expression of disulfide bond-dependent proteins in *Escherichia coli*. *Microbial Cell Factories*, *8*(1), 26. http://doi.org/10.1186/1475-2859-8-26.

de Massy, B., Studier, F. W., Dorgai, L., Appelbaum, E., & Weisberg, R. A. (1984). Enzymes and sites of genetic recombination: studies with gene-3 endonuclease of phage T7 and with site-affinity mutants of phage lambda. *Cold Spring Harbor Symposia on Quantitative Biology*, *49*, 715–26.

Derman, A., Prinz, W., Belin, D., & Beckwith, J. (1993). Mutations that allow disulfide bond formation in the cytoplasm of *Escherichia coli*. *Science*, *262*(5140), 1744–1747.

Desai, N. A., & Shankar, V. (2003). Single-strand-specific nucleases. *FEMS Microbiology Reviews*, *26*(5), 457–491. http://doi.org/10.1111/j.1574-6976.2003.tb00626.x.

di Guana, C., Lib, P., Riggsa, P. D., Hiroshi, I. (1988). Vectors that facilitate the expression and purification of foreign peptides in *Escherichia coli* by fusion to maltose-binding protein. *Gene*, *67*(1), 21–30.

Dieckman, L., Gu, M., Stols, L., Donnelly, M. I., & Collart, F. R. (2002). High throughput methods for gene cloning and expression. *Protein Expression and Purification*, *25*(1), 1–7. http://doi.org/10.1006/prep.2001.1602.

Dong, H., Nilsson, L., & Kurland, C. G. (1995). Gratuitous overexpression of genes in *Escherichia coli* leads to growth inhibition and ribosome destruction. *Journal of Bacteriology*, *177*(6), 1497–504.

Dougherty, W. G., Carrington, J. C., Cary, S. M., & Parks, T. D. (1988). Biochemical and mutational analysis of a plant virus polyprotein cleavage site. *The EMBO Journal*, *7*(5), 1281–1287.

Dutertre, S., & Lewis, R. J. (2010). Use of Venom Peptides to Probe Ion Channel Structure and Function. *Journal of Biological Chemistry*, *285*(18), 13315–13320. http://doi.org/10.1074/jbc.R109.076596.

Dutertre, S., Undheim, E. A. B., Pineda, S. S., Jin, A.-H., Lavergne, V., Fry, B. G., … King, G. F. (2015). Venoms-Based Drug Discovery : Proteomic and Transcriptomic Approaches. *RSC Drug Discovery*, (42), 80–96.

Dyson, M. R., Shadbolt, S. P., Vincent, K. J., Perera, R. L., & McCafferty, J. (2004). Production of soluble mammalian proteins in *Escherichia coli*: identification of protein features that

correlate with successful expression. *BMC Biotechnology*, *4*, 32. http://doi.org/10.1186/1472-6750-4-32.

Edge, M. D., Green, A. R., Heathclife, G. R., Meacock, P. A., Schuch, W., Scanlon, D. B., … Markham, A. F. (1981). Total synthesis of a human leukocyte interferon gene. *Nature*, *292*, 756–762.

Eisenberg, D., Marcotte, E. M., Xenarios, I., & Yeates, T. O. (2000). Protein function in the post-genomic era. *Nature*, *405*(6788), 823–826. http://doi.org/10.1038/35015694.

Englander, S. W. (2000). Protein folding intermediates and pathways studied by hydrogen exchange. *Annual Review of Biophysics and Biomolecular Structure*, *29*, 213–238.

Eren, M., & Swenson, R. P. (1989). Chemical synthesis and expression of a synthetic gene for the flavodoxin from Clostridium MP. *The Journal of Biological Chemistry*, *264*(25), 14874–9.

Escoubas, P., Bernard, C., Lambeau, G., Lazdunski, M., & Darbon, H. (2003). Recombinant production and solution structure of PcTx1 , the specific peptide inhibitor of ASIC1a proton-gated cation channels. *Protein Science*, *12*, 1332–1343. http://doi.org/10.1110/ps.0307003.nels.

Escoubas, P., & King, G. F. (2009). Venomics as a drug discovery platform. *Expert Review of Proteomics*, *6*(3), 221–224. http://doi.org/10.1586/epr.09.45.

Escoubas, P., Quinton, L., & Nicholson, G. M. (2008). Venomics: unravelling the complexity of animal venoms with mass spectrometry. *Journal of Mass Spectrometry : JMS*, *43*(7), 279–295. http://doi.org/10.1002/jms.

Escoubas, P., Sollod, B., & King, G. F. (2006). Venom landscapes: Mining the complexity of spider venoms via a combined cDNA and mass spectrometric approach. *Toxicon*, *47*(6), 650–663. http://doi.org/10.1016/j.toxicon.2006.01.018.

Fernandes-Pedrosa, M. F., Félix-Silva, J., & Menezes, Y. a S. (2013). Toxins from Venomous Animals: Gene Cloning , Protein Expression and Biotechnological Applications. *An Integrated View of the Molecular Recognition and Toxinologu - From Analytical Procedures to Biomedical Applications*, *2*, 23–72. http://doi.org/10.5772/52380.

Fuglsang, A. (2003). The effective number of codons for individual amino acids: Some codons are more optimal than others. *Gene*, *320*(1-2), 185–190. http://doi.org/10.1016/S0378-1119(03)00829-1.

Fuhrmann, M., Oertel, W., Berthold, P., & Hegemann, P. (2005). Removal of mismatched bases from synthetic genes by enzymatic mismatch cleavage. *Nucleic Acids Research*, *33*(6), e58. http://doi.org/10.1093/nar/gni058.

Gaj, T., Gersbach, C. A., & Barbas, C. F. (2013). ZFN, TALEN, and CRISPR/Cas-based methods for genome engineering. *Trends in Biotechnology*, *31*(7), 397–405. http://doi.org/10.1016/j.tibtech.2013.04.004.

Gao, X., Gulari, E., & Zhou, X. (2004). In situ synthesis of oligonucleotide microarrays. *Biopolymers*, *73*(5), 579–96. http://doi.org/10.1002/bip.20005.

Gao, X., Yo, P., Keith, A., Ragan, T. J., & Harris, T. K. (2003). Thermodynamically balanced inside-out (TBIO) PCR-based gene synthesis: a novel method of primer design for high-fidelity assembly of longer gene sequences. *Nucleic Acids Research*, *31*(22), e143.

Geertsma, E. R., & Dutzler, R. (2011). A versatile and efficient high-throughput cloning tool for structural biology. *Biochemistry*, *50*(15), 3272–8. http://doi.org/10.1021/bi200178z.

Gibson, D. G., Benders, G. A., Andrews-Pfannkoch, C., Denisova, E. A., Baden-Tillson, H., Zaveri, J., … Smith, H. O. (2008). Complete chemical synthesis, assembly, and cloning of a Mycoplasma genitalium genome. *Science (New York, N.Y.)*, *319*(5867), 1215–20. http://doi.org/10.1126/science.1151721.

Gilles, N., & Servent, D. (2014). The European FP7 Venomics Project. *Future Medicinal Chemistry*, *6*(15), 1611–2. http://doi.org/10.4155/fmc.14.85.

Gordeeva, T. L., Borschevskaya, L. N., & Sineoky, S. P. (2010). Improved PCR-based gene synthesis method and its application to the Citrobacter freundii phytase gene codon modification. *Journal of Microbiological Methods*, *81*(2), 147–152. http://doi.org/10.1016/j.mimet.2010.02.013.

Grundström, T., Zenke, W. M., Wintzerith, M., Matthes, H. W., Staub, A., & Chambon, P. (1985). Oligonucleotide-directed mutagenesis by microscale “shot-gun” gene synthesis. *Nucleic Acids Research*, *13*(9), 3305–16.

Gustafsson, C., Govindarajan, S., & Minshull, J. (2004). Codon bias and heterologous protein expression. *Trends in Biotechnology*, *22*(7), 346–353. http://doi.org/10.1016/j.tibtech.2004.04.006.

Haag, A. F., & Ostermeier, C. (2009). Positive-selection vector for direct protein expression. *BioTechniques*, *46*(6), 453–7. http://doi.org/10.2144/000113091.

Hammarström, M., Hellgren, N., van den Berg, S., Berglund, H., & Härd, T. (2002). Rapid screening for improved solubility of small human proteins produced as fusion proteins in *Escherichia coli*. *Protein Science*, *11*, 313–321. http://doi.org/10.1110/ps.22102.biology.

Hartley, J. L. (2006). Cloning technologies for protein expression and purification. *Current Opinion in Biotechnology*, *17*(4), 359–66. http://doi.org/10.1016/j.copbio.2006.06.011.

Hartley, J. L., Temple, G. F., & Brasch, M. A. (2000a). DNA Cloning Using In Vitro Site-Specific Recombination. *Genome Research*, *10*(11), 1788–1795. http://doi.org/10.1101/gr.143000.

Hartley, J. L., Temple, G. F., & Brasch, M. A. (2000b). DNA cloning using in vitro site-specific recombination. *Genome Research*, *10*(11), 1788–95.

Harvey, A. L. (2014). Toxins and drug discovery. *Toxicon*, *92*, 193–200. http://doi.org/10.1016/j.toxicon.2014.10.020.

Hayashi, N., Welschof, M., Zewe, M., Braunagel, M., Dübel, S., Breitling, F., & Little, M. (1994). Simultaneous mutagenesis of antibody CDR regions by overlap extension and PCR. *BioTechniques*, *17*(2), 310, 312, 314–5.

He, Q., Han, W., Huo, L., Zhang, J., Lin, Y., Chen, P., & Liang, S. (2010). ATDB 2.0: A database integrated toxin-ion channel interaction data. *Toxicon*, *56*, 644–647.

Henaut, A., & Danchin, A. (1996). Analysis and predictions from *Escherichia coli* sequences. *In: Neidhart FC, Curtiss RI, Ingraham J, Lin E, Brooks Low K, et Al., Eds. Escherichia coli and Salmonella Typhimurium Cellular and Molecular Biology. Washington, D. C.: ASM Press*, 2047–2006.

Hobom, B. (1980). Surgery of genes. At the doorstep of synthetic biology. *Medizin. Klinik*, *75*, 14–21.

Hoover, D. M., & Lubkowski, J. (2002). DNAWorks: an automated method for designing oligonucleotides for PCR-based gene synthesis. *Nucleic Acids Research*, *30*(10), e43.

Horton, R. M., Hunt, H. D., Ho, S. N., Pullen, J. K., & Pease, L. R. (1989). Engineering hybrid genes without the use of restriction enzymes: gene splicing by overlap extension. *Gene*, *77*(1), 61–8.

Horvath, S. J., Firca, J. R., Hunkapiller, T., Hunkapiller, M. W., & Hood, L. (1987). An automated DNA synthesizer employing deoxynucleoside 3’-phosphoramidites. *Methods in Enzymology*, *154*, 314–26.

Hu, L.-L., Zhang, S.-S., Li, X.-X., & Wang, B.-L. (2010). The use of the ccdB lethal gene for constructing a zero background vector in order to clone blunt-end PCR Products. *Molecular Biology*, *44*(1), 161–164. http://doi.org/10.1134/S0026893310010206.

Huang, M. C., Cheong, W. C., Lim, L. S., & Li, M.-H. (2012). A simple, high sensitivity mutation screening using Ampligase mediated T7 endonuclease I and Surveyor nuclease with microfluidic capillary electrophoresis. *Electrophoresis*, *33*(5), 788–796. http://doi.org/10.1002/elps.201100460.

Hughes, R. a., Miklos, A. E., & Ellington, A. D. (2011). *Gene Synthesis: Methods and Applications* (Vol. 498). http://doi.org/10.1016/B978-0-12-385120-8.00012-7.

Hwang, P. M., Pan, J. S., & Sykes, B. D. (2014). Targeted expression, purification, and cleavage of fusion proteins from inclusion bodies in *Escherichia coli*. *FEBS Letters*, *588*(2), 247–52. http://doi.org/10.1016/j.febslet.2013.09.028.

Jenny, R. J., Mann, K. G., & Lundblad, R. L. (2003). A critical review of the methods for cleavage of fusion proteins with thrombin and factor Xa. *Protein Expression and Purification*, *31*(1), 1–11. http://doi.org/10.1016/S1046-5928(03)00168-2.

Juarez, P., Sanz, L., & Calvete, J. J. (2004). Snake venomics: characterization of protein families in Sistrurus barbouri venom by cysteine mapping, N-terminal sequencing, and tandem mass spectrometry analysis. *Proteomics 4*, 327–338.

Kaas, Q., Westermann, J., Halai, R., Wang, C., & Craik, D. (2008). ConoServer, a database for conopeptide sequences and structures. *Bioinformatics*, *24*(3), 445–6.

Kapust, R. B., Tözsér, J., Copeland, T. D., & Waugh, D. S. (2002). The P1' specificity of tobacco etch virus protease. *Biochemical and Biophysical Research Communications*, *294*(5), 949–55. http://doi.org/10.1016/S0006-291X(02)00574-0.

Kapust, R. B., & Waugh, D. S. (1999). *Escherichia coli* maltose-binding protein is uncommonly effective at promoting the solubility of polypeptides to which it is fused. *Protein Science : A Publication of the Protein Society*, *8*(8), 1668–74. http://doi.org/10.1110/ps.8.8.1668.

Karbat, I., Turkov, M., Cohen, L., Kahn, R., Gordon, D., Gurevitz, M., & Frolow, F. (2007). X-ray structure and mutagenesis of the scorpion depressant toxin LqhIT2 reveals key determinants crucial for activity and anti-insect selectivity. *Journal of Molecular Biology*, *366*(2), 586–601. http://doi.org/10.1016/j.jmb.2006.10.085.

King, G. F. (2011). Venoms as a platform for human drugs: translating toxins into therapeutics. *Expert Opinion on Biological Therapy*, *11*(11), 1469–1484. http://doi.org/10.1517/14712598.2011.621940.

Klint, J. K., Senff, S., Saez, N. J., Seshadri, R., Lau, H. Y., Bende, N. S., … King, G. F. (2013). Production of Recombinant Disulfide-Rich Venom Peptides for Structural and Functional Analysis via Expression in the Periplasm of *E. coli*. *PLoS ONE*, *8*(5), e63865. http://doi.org/10.1371/journal.pone.0063865.

Klint, J. K., Smith, J. J., Vetter, I., Rupasinghe, D. B., Er, S. Y., Senff, S., … King, G. F. (2015). Seven novel modulators of the analgesic target NaV1.7 uncovered using a high-throughput venom-based discovery approach. *British Journal of Pharmacology*, *172*(10), 2445–2458. http://doi.org/10.1111/bph.13081.

Kodumal, S. J., Patel, K. G., Reid, R., Menzella, H. G., Welch, M., & Santi, D. V. (2004). Total synthesis of long DNA sequences: synthesis of a contiguous 32-kb polyketide synthase gene cluster. *Proceedings of the National Academy of Sciences of the United States of America*, *101*(44), 15573–8. http://doi.org/10.1073/pnas.0406911101.

Koehn, J., & Hunt, I. (2009). High-Throughput Protein Production (HTPP): A Review of Enabling Technologies to Expedite Protein Production. In S. A. Doyle (Ed.), *Methods in Molecular Biology: High Throughput Protein Expression and Purification, vol. 498* (Vol. 498, pp. 1–18). Totowa, NJ: Humana Press. http://doi.org/10.1007/978-1-59745-196-3.

Kosuri, S., & Church, G. M. (2014). Large-scale de novo DNA synthesis: technologies and applications. *Nature Methods*, *11*(5), 499–507. http://doi.org/10.1038/nmeth.2918.

Kudla, G., Murray, A. W., Tollervey, D., & Plotkin, J. B. (2009). Coding-sequence determinants of gene expression in *Escherichia coli*. *Science (New York, N.Y.)*, *324*(5924), 255–8. http://doi.org/10.1126/science.1170160.

Lashkari, D. A., Hunicke-Smith, S. P., Norgren, R. M., Davis, R. W., & Brennan, T. (1995). An automated multiplex oligonucleotide synthesizer: development of high-throughput, low-cost DNA synthesis. *Proceedings of the National Academy of Sciences of the United States of America*, *92*(17), 7912–5.

LaVallie, E. R., DiBlasio, E. A., Kovacic, S., Grant K. L., Schendel, P. F., McCoy, J. M. (1993). A thioredoxin gene fusion expression system that circumvents inclusion body formation in the *E. coli* cytoplasm. *Biotechnology (N. Y.)*, *11*(2), 187–193.

Lavergne, V., Alewood, P. F., Mobli, M., & King, G. F. (2015). The Structural Universe of Disulfide-Rich Venom Peptides. *RSC Drug Discovery*, (42), 37.

LeProust, E. M., Peck, B. J., Spirin, K., McCuen, H. B., Moore, B., Namsaraev, E., & Caruthers, M. H. (2010). Synthesis of high-quality libraries of long (150mer) oligonucleotides by a novel depurination controlled process. *Nucleic Acids Research*, *38*(8), 2522–40. http://doi.org/10.1093/nar/gkq163.

Lesley, S. a. (2001). High-throughput proteomics: protein expression and purification in the postgenomic world. *Protein Expression and Purification*, *22*(2), 159–64. http://doi.org/10.1006/prep.2001.1465.

Lewis, R. J., & Garcia, M. L. (2003). Therapeutic potential of venom peptides. *Nature Reviews Drug Discovery*, *2*(10), 790–802. http://doi.org/10.1038/nrd1197.

Li, M. Z., & Elledge, S. J. (2005). MAGIC, an in vivo genetic method for the rapid construction of recombinant DNA molecules. *Nature Genetics*, *37*, 311–319.

Li, M. Z., & Elledge, S. J. (2007). Harnessing homologous recombination in vitro to generate recombinant DNA via SLIC. *Nature Methods*, *4*, 251–256.

Lithwick, G., & Margalit, H. (2003). Hierarchy of sequence-dependent features associated with prokaryotic translation. *Genome Research*, *13*(12), 2665–73. http://doi.org/10.1101/gr.1485203.

Ma, S., Saaem, I., & Tian, J. (2012). Error correction in gene synthesis technology. *Trends in Biotechnology*, *30*(3), 147–154. http://doi.org/10.1016/j.tibtech.2011.10.002.

Macauley-Patrick, S., Fazenda, M. L., McNeil, B., & Harvey, L. M. (2005). Heterologous protein production using the Pichia pastoris expression system. *Yeast*, *22*(4), 249–270. http://doi.org/10.1002/yea.1208.

Mamelak, A., & Jacoby, D. (2007). Targeted delivery of antitumoral therapy to glioma and other malignancies with synthetic chlorotoxin (TM-601). *Expert Opin. Drug Deliv.*, *4*(2), 175–186.

Mancia, F., & Love, J. (2011). High throughput platforms for structural genomics of integral membrane proteins. *Current Opinion in Structural Biology*, *21*(4), 517–22. http://doi.org/10.1016/j.sbi.2011.07.001.

Marsischky, G., & LaBaer, J. (2004). Many paths to many clones: a comparative look at high-throughput cloning methods. *Genome Research*, *14*(10B), 2020–8. http://doi.org/10.1101/gr.2528804.

Meng, E., Cai, T. F., Li, W. Y., Zhang, H., Liu, Y. B., Peng, K., … Zhang, D. Y. (2011). Functional expression of spider neurotoxic peptide Huwentoxin-I in *E. coli*. *PLoS ONE*, *6*(6), 3–8. http://doi.org/10.1371/journal.pone.0021608.

Miljanich, G. (2004). Ziconotide: neuronal calcium channel blocker for treating severe chronic pain. *Curr. Med. Chem.*, *11*(23), 3029–3040.

Moffatt, B. A., &, & Studier, F. W. (1987). T7 lysozyme inhibits transcription by T7 RNA polymerase. *Cell*, *49*(2), 221–227.

Mouhat, S., Jouirou, B., Mosbah, A., De Waard, M., & Sabatier, J.-M. (2004). Diversity of folds in animal toxins acting on ion channels. *The Biochemical Journal*, *378*(Pt 3), 717–726. http://doi.org/10.1042/BJ20031860.

Näreoja, K., & Näsman, J. (2012). Selective targeting of G-protein-coupled receptor subtypes with venom peptides. *Acta Physiologica*, *204*(2), 186–201. http://doi.org/10.1111/j.1748-1716.2011.02305.x.

Nikaido, H. (1994). Maltose transport system of *Escherichia coli*: an ABC-type transporter. *FEBS Letters*, *346*, 55–58.

Nozach, H., Fruchart-Gaillard, C., Fenaille, F., Beau, F., Ramos, O. H. P., Douzi, B., … Dive, V. (2013). High throughput screening identifies disulfide isomerase DsbC as a very efficient partner for recombinant expression of small disulfide-rich proteins in *E. coli*. *Microbial Cell Factories*, *12*(1), 37. http://doi.org/10.1186/1475-2859-12-37.

O'Reilly, A. O., Cole, A. R., Lopes, J. L. S., Lampert, A., & Wallace, B. A. (2014). Chaperone-mediated native folding of a β-scorpion toxin in the periplasm of *Escherichia coli*. *Biochimica et Biophysica Acta*, *1840*(1), 10–5. http://doi.org/10.1016/j.bbagen.2013.08.021.

Parks, T. D., Leuther, K. K., Howard, E. D., Johnston, S. A., & Dougherty, W. G. (1994). Release of proteins and peptides from fusion proteins using a recombinant plant virus proteinase. *Analytical Biochemistry*. http://doi.org/10.1006/abio.1994.1060.

Peleg, Y., & Unger, T. (2012). Chemical Genomics and Proteomics. In E. D. Zanders (Ed.), *Chemical Genomics and Proteomics: Reviews and Protocols, Methods in Molecular Biology, vol. 800* (Vol. 800, pp. 173–186). Totowa, NJ: Humana Press. http://doi.org/10.1007/978-1-61779-349-3.

Polizzi, K. M. (2013). *Synthetic Biology*. http://doi.org/10.1007/978-1-62703-625-2.

Prashanth, J. R., Lewis, R. J., & Dutertre, S. (2012). Towards an integrated venomics approach for accelerated conopeptide discovery. *Toxicon*, *60*, 470–477. http://doi.org/10.1016/j.toxicon.2012.04.340.

Quan, J., & Tian, J. (2009). Circular polymerase extension cloning of complex gene libraries and pathways. *PloS One*, *4*(7), e6441. http://doi.org/10.1371/journal.pone.0006441.

Quax, T. E. F., Claassens, N. J., Söll, D., & van der Oost, J. (2015). Codon Bias as a Means to Fine-Tune Gene Expression. *Molecular Cell*, *59*(2), 149–161. http://doi.org/10.1016/j.molcel.2015.05.035.

Quentin, K., & Craik, D. J. (2015). Bioinformatics-Aided Venomics. *Toxins*, *7*, 2159–2187. http://doi.org/10.3390/toxins7062159.

Raran-Kurussi, S., & Waugh, D. S. (2012). The Ability to Enhance the Solubility of Its Fusion Partners Is an Intrinsic Property of Maltose-Binding Protein but Their Folding Is Either Spontaneous or Chaperone-Mediated. *PLoS ONE*, *7*(11). http://doi.org/10.1371/journal.pone.0049589.

Rosano, G. L., & Ceccarelli, E. a. (2014). Recombinant protein expression in *Escherichia coli*: Advances and challenges. *Frontiers in Microbiology*, *5*(April), 1–17. http://doi.org/10.3389/fmicb.2014.00172.

Saaem, I., Ma, S., Quan, J., & Tian, J. (2012). Error correction of microchip synthesized genes using Surveyor nuclease. *Nucleic Acids Research*, *40*(3), 1–8. http://doi.org/10.1093/nar/gkr887.

Saez, N. J., Mobli, M., Bieri, M., Chassagnon, I. R., Malde, A. K., Gamsjaeger, R., … King, G. F. (2011). A dynamic pharmacophore drives the interaction between Psalmotoxin-1 and the putative drug target acid-sensing ion channel 1a. *Mol Pharmacol*, *80*(5), 796–808. http://doi.org/10.1124/mol.111.072207.

Saez, N. J., Nozach, H., Blemont, M., & Vincentelli, R. (2014). *High Throughput Quantitative Expression Screening and Purification Applied to Recombinant Disulfide-rich Venom Proteins Produced in E. coli*. *Jove-Journal of Visualized Experiments* (pp. 1–15). http://doi.org/10.3791/51464.

Saez, N. J., & Vincentelli, R. (2014). High-Throughput Expression Screening and Purification of Recombinant Proteins in *E. coli*. *Methods in Molecular Biology (Clifton, NJ)*, *1091*, 33–53.

Saiki, R. K., Scharf, S., Faloona, F., Mullis, K. B., Horn, G. T., Erlich, H. A., & Arnheim, N. (1985). Enzymatic amplification of beta-globin genomic sequences and restriction site analysis for diagnosis of sickle cell anemia. *Science*, *230*(4732), 1350–1354.

Salinas, G., Pellizza, L., Margenat, M., Fló, M., & Fernández, C. (2011). Tuned *Escherichia*

coli as a host for the expression of disulfide-rich proteins. *Biotechnology Journal*, *6*(6), 686–99. http://doi.org/10.1002/biot.201000335.

Sandhu, G. S., Aleff, R. A., & Kline, B. C. (1992). Dual asymmetric PCR: one-step construction of synthetic genes. *BioTechniques*, *12*(1), 14–6.

SAS. (2004). *SAS User´s Guide: Statistics*. SAS Institute In Cary NC.

Satakarni, M., & Curtis, R. (2011). Production of recombinant peptides as fusions with SUMO. *Protein Expression and Purification*, *78*(2), 113–9. http://doi.org/10.1016/j.pep.2011.04.015.

Scheich, C., Sievert, V., & Büssow, K. (2003). An automated method for high-throughput protein purification applied to a comparison of His-tag and GST-tag affinity chromatography. *BMC Biotechnology*, *3*, 12. http://doi.org/10.1186/1472-6750-3-12.

Schwartz, J. J., Lee, C., & Shendure, J. (2012). Accurate gene synthesis with tag-directed retrieval of sequence-verified DNA molecules. *Nature Methods*, *9*(9), 913–5. http://doi.org/10.1038/nmeth.2137.

Sharp, P. M., & Li, W. H. (1987). The codon adaptation index-a measure of directional synonymous codon usage bias, and its potential applications. *Nucleic Acids Research*, *15*(3), 1281–1295. http://doi.org/10.1093/nar/15.3.1281.

Shih, Y., Kung, W., Chen, J., Yeh, C., Wang, A. H., & Wang, T. (2002). High-throughput screening of soluble recombinant proteins. *Protein Science*, *11*(7), 1714–1719. http://doi.org/10.1110/ps.0205202.Edwards.

Sierzchala, A. B., Dellinger, D. J., Betley, J. R., Wyrzykiewicz, T, K., Yamada, C. M., & Caruthers, M. H. (2003). Solid-Phase Oligodeoxynucleotide Synthesis: A Two-Step Cycle Using Peroxy Anion Deprotection. *J. Am. Chem. Soc.*, *125*, 13427–13441.

Silber, J. R., & Loeb, L. A. (1981). S1 nuclease does not cleave DNA at single-base mis-matches. *Biochimica et Biophysica Acta*, *656*(2), 256–64.

Sindelar, L. E., & Jaklevic, J. M. (1995). High-throughput DNA synthesis in a multichannel format. *Nucleic Acids Research*, *23*(6), 982–7.

Smith, D. B, & Johnson, K. S. (1988). Single-step purification of polypeptides expressed in *Escherichia coli* as fusions with glutathione S-transferase. *Gene*, *67*(1), 31–40.

Smith, J., & Modrich, P. (1997). Removal of polymerase-produced mutant sequences from PCR products. *Proceedings of the National Academy of Sciences of the United States of America*, *94*(13), 6847–6850. http://doi.org/10.1073/pnas.94.13.6847.

Staden, R. (1979). A strategy of DNA sequencing employing computer programs. *Nucleic Acids Research*, *6*(7), 2601–10. http://doi.org/10.1093/nar/6.7.2601.

Stemmer, W. P., Crameri, A., Ha, K. D., Brennan, T. M., & Heyneker, H. L. (1995). Single-step assembly of a gene and entire plasmid from large numbers of oligodeoxyribonucleotides. *Gene*, *164*(1), 49–53.

Strizhov, N., Keller, M., Mathur, J., Koncz-Kálmán, Z., Bosch, D., Prudovsky, E., … Zilberstein, A. (1996). A synthetic cryIC gene, encoding a Bacillus thuringiensis delta-endotoxin, confers Spodoptera resistance in alfalfa and tobacco. *Proceedings of the National Academy of Sciences of the United States of America*, *93*(26), 15012–7.

Studier, F. W. (2005a). Protein production by auto-induction in high density shaking cultures. *Protein Expression and Purification*, *41*(1), 207–34.

Studier, F. W. (2005b). Protein production by auto-induction in high-density shaking cultures. *Protein Expression and Purification*, *41*(1), 207–234. http://doi.org/10.1016/j.pep.2005.01.016.

Tachibana, A., Tohiguchi, K., Ueno, T., Setogawa, Y., Harada, A., & Tanabe, T. (2009). Preparation of long sticky ends for universal ligation-independent cloning: sequential T4 DNA polymerase treatments. *Journal of Bioscience and Bioengineering*, *107*(6), 668–9. http://doi.org/10.1016/j.jbiosc.2009.01.019.

Takacs, Z., & York, N. (2014). *Animal Venoms in Medicine*. *Encyclopedia of Toxicology* (Third Edit, Vol. 1). Elsevier. http://doi.org/10.1016/B978-0-12-386454-3.01241-0.

Takagi, M., Nishioka, M., Kakihara, H., Kitabayashi, M., Inoue, H., Kawakami, B., … Imanaka, T. (1997). Characterization of DNA polymerase from Pyrococcus sp. strain KOD1 and its application to PCR. *Applied and Environmental Microbiology*, *63*(11), 4504–10.

Terlau, H., & Olivera, B. M. (2004). Conus Venoms: A rich Source of Novel Ion Channel-Targeted Peptides. *American Physiological Society*.

Terpe, K. (2003). Overview of tag protein fusions: from molecular and biochemical fundamentals to commercial systems. *Applied Microbiology and Biotechnology*, *60*(5), 523–33. http://doi.org/10.1007/s00253-002-1158-6.

Tian, J., Ma, K., & Saaem, I. (2009). Advancing high-throughput gene synthesis technology. *Molecular BioSystems*, *5*(7), 714. http://doi.org/10.1039/b822268c.

Till, B. J., Burtner, C., Comai, L., & Henikoff, S. (2004). Mismatch cleavage by single-strand specific nucleases. *Nucleic Acids Research*, *32*(8), 2632–2641. http://doi.org/10.1093/nar/gkh599.

Tindall, K. R., & Kunkel, T. A. (1988). Fidelity of DNA synthesis by the Thermus aquaticus DNA polymerase. *Biochemistry*, *27*(16), 6008–13.

Trim, S. A., & Trim, C. M. (2013). Venom: the sharp end of pain therapeutics. *British Journal of Pain*, *7*(4), 179–188. http://doi.org/10.1177/2049463713502005.

Tsuji, T., & Niida, Y. (2008). Development of a simple and highly sensitive mutation screening system by enzyme mismatch cleavage with optimized conditions for standard laboratories. *Electrophoresis*, *29*(7), 1473–1483. http://doi.org/10.1002/elps.200700729.

Uetz, P., & Hošek, J. (1996). The Reptile Database. Retrieved January 25, 2016, from www.reptile-database.org.

Van Den Berg, S., Löfdahl, P. Å., Härd, T., & Berglund, H. (2006). Improved solubility of TEV protease by directed evolution. *Journal of Biotechnology*, *121*(3), 291–298. http://doi.org/10.1016/j.jbiotec.2005.08.006.

Venomics, C. (2012). Venomics project. Retrieved February 4, 2016, from http://www.venomics-project.eu/.

Vetter, I., Davis, J. L., Rash, L. D., Anangi, R., Mobli, M., Alewood, P. F., … King, G. F. (2011). Venomics: a new paradigm for natural products-based drug discovery. *Amino Acids*, *40*(1), 15–28. http://doi.org/10.1007/s00726-010-0516-4.

Vincentelli, R., Cimino, A., Geerlof, A., Kubo, A., Satou, Y., & Cambillau, C. (2011). High-throughput protein expression screening and purification in *Escherichia coli*. *Methods*, *55*(1), 65–72. http://doi.org/10.1016/j.ymeth.2011.08.010.

Wan, W., Li, L., Xu, Q., Wang, Z., Yao, Y., Wang, R., … Hong, J. (2014). Error removal in microchip-synthesized DNA using immobilized MutS. *Nucleic Acids Research*, *42*(12), e102. http://doi.org/10.1093/nar/gku405.

Way, J. C., Collins, J. J., Keasling, J. D., & Silver, P. A. (2014). Integrating Biological Redesign: Where Synthetic Biology Came From and Where It Needs to Go. *Cell*, *157*(1), 151–161. http://doi.org/10.1016/j.cell.2014.02.039.

Weber, J. L., & Myers, E. W. (1997). Human whole-genome shotgun sequencing. *Genome Research*, *7*(715), 401–409. http://doi.org/10.1101/gr.7.5.401.

Weinstock, G. M., Robinson, G. E., Gibbs, R. A., Weinstock, G. M., Weinstock, G. M., Robinson, G. E., … Wright, R. (2006). Insights into social insects from the genome of the honeybee Apis mellifera. *Nature*, *443*(7114), 931–949. http://doi.org/10.1038/nature05260.

Welch, M., Govindarajan, S., Ness, J. E., Villalobos, A., Gurney, A., Minshull, J., & Gustafsson, C. (2009). Design Parameters to Control Synthetic Gene Expression in *Escherichia coli*. *PLoS ONE*, *4*(9), e7002. http://doi.org/10.1371/journal.pone.0007002.

Welch, M., Villalobos, A., Gustafsson, C., & Minshull, J. (2009). You're one in a googol: optimizing genes for protein expression. *Journal of The Royal Society Interface*, *6*(Suppl_4), S467–S476. http://doi.org/10.1098/rsif.2008.0520.focus.

Welch, M., Villalobos, A., Gustafsson, C., & Minshull, J. (2011). *Designing Genes for Successful Protein Expression*. *Synthetic Biology Part B* (1st ed., Vol. 498). Elsevier Inc. http://doi.org/10.1016/B978-0-12-385120-8.00003-6.

White, M. F., Giraud-Panis, M. J., Pöhler, J. R., & Lilley, D. M. (1997). Recognition and manipulation of branched DNA structure by junction-resolving enzymes. *Journal of Molecular Biology*, *269*(5), 647–64. http://doi.org/10.1006/jmbi.1997.1097.

Wild, J., Hradecna, Z., & Szybalski, W. (2002). Conditionally Amplifiable BACs : Switching From Single-Copy to High-Copy Vectors and Genomic C.lones. *Genome*, *12*(1), 1434–1444. http://doi.org/10.1101/gr.130502.replication

World Spider Catalog. (2016). Retrieved May 4, 2016, from http://www.wsc.nmbe.ch/.

Wu, G., Wolf, J. B., Ibrahim, A. F., Vadasz, S., Gunasinghe, M., & Freeland, S. J. (2006). Simplified gene synthesis: a one-step approach to PCR-based gene construction. *Journal of Biotechnology*, *124*(3), 496–503. http://doi.org/10.1016/j.jbiotec.2006.01.015.

Wu, J. J., He, L. L., Zhou, Z., & Chi, C. W. (2002). Gene expression, mutation, and structure-function relationship of scorpion toxin BmP05 active on SKCa channels. *Biochemistry*, *41*(8), 2844–2849. http://doi.org/10.1021/bi011367z.

Wurm, Y., Wang, J., Riba-Grognuz, O., Corona, M., Nygaard, S., Hunt, B. G., … Keller, L. (2011). The genome of the fire ant Solenopsis invicta. *Proceedings of the National Academy of Sciences of the United States of America*, *108*(14), 5679–5684. http://doi.org/10.1073/pnas.1009690108.

Xiong, A.-S., Peng, R.-H., Zhuang, J., Gao, F., Li, Y., Cheng, Z.-M., & Yao, Q.-H. (2008). Chemical gene synthesis: strategies, softwares, error corrections, and applications. *FEMS Microbiology Reviews*, *32*(3), 522–40. http://doi.org/10.1111/j.1574-6976.2008.00109.x.

Xiong, A.-S., Yao, Q.-H., Peng, R.-H., Duan, H., Li, X., Fan, H.-Q., … Li, Y. (2006). PCR-based accurate synthesis of long DNA sequences. *Nature Protocols*, *1*(2), 791–797. http://doi.org/10.1038/nprot.2006.103.

Xiong, A.-S., Yao, Q.-H., Peng, R.-H., Li, X., Fan, H.-Q., Cheng, Z.-M., & Li, Y. (2004). A simple, rapid, high-fidelity and cost-effective PCR-based two-step DNA synthesis method for long gene sequences. *Nucleic Acids Research*, *32*(12), e98–e98. http://doi.org/10.1093/nar/gnh094.

Yang, B., Wen, X., Kodali, N. S., Oleykowski, C. a., Miller, C. G., Kulinski, J., … Yeung, A. T. (2000). Purification , Cloning , and Characterization of the CEL I Nuclease †. *Biochemistry*, *39*(13), 3533–3541. http://doi.org/10.1021/bi992376z.

Yang, J., Zhang, Z., Zhang, X. A., & Luo, Q. (2010). A ligation-independent cloning method using nicking DNA endonuclease. *BioTechniques*, *49*, 817–821. http://doi.org/10.2144/000113520.

Yang, S., Xiao, Y., Kang, D., Liu, J., Li, Y., Undheim, E. a B., … King, G. F. (2013). Discovery of a selective NaV1.7 inhibitor from centipede venom with analgesic efficacy exceeding morphine in rodent pain models. *Proceedings of the National Academy of Sciences of the United States of America*, *110*(43), 17534–9. http://doi.org/10.1073/pnas.1306285110.

Yeung, A. T., Hattangadi, D., Blakesley, L., & Nicolas, E. (2005). Enzymatic mutation detection technologies. *BioTechniques*, *38*(5), 749–58.

Young, L., & Dong, Q. (2004). Two-step total gene synthesis method. *Nucleic Acids Research*, *32*(7), e59. http://doi.org/10.1093/nar/gnh058.

Zampini, M., Stevens, P. R., Pachebat, J. a., Kingston-Smith, A., Mur, L. a. J., & Hayes, F. (2015). RapGene: a fast and accurate strategy for synthetic gene assembly in *Escherichia coli*. *Scientific Reports*, *5*, 11302. http://doi.org/10.1038/srep11302.

Zhong, Y., Song, B., Mo, G., Yuan, M., Li, H., Wang, P., … Lu, Q. (2014). A novel neurotoxin from venom of the spider, Brachypelma albopilosum. *PloS One*, *9*(10), e110221. http://doi.org/10.1371/journal.pone.0110221.

ANNEXES

Supplemental information – Chapter 3

Table S3. 1| Sequences properties of genes A, B and C used to optimize the PCR conditions.

Gene name	Primary protein sequence	Protein size (aa)	Gene sequence	Gene size (nt)	GC content (%)	CAI	Recombinant gene size (nt), PCA-DT	Recombinant gene size (nt), PCA-DTF
A	IRCFITPDITSK DCPNGHVCYT KTWCDAFCSI RGKRVDLGCA ATCPTVKTGV DIQCCSTDNC NPFPTRKRP	71	ATCCGCTGCTTCATCACG CCGGACATTACCTCAAAA GATTGCCCGAACGGTCA CGTCTGCTACACGAAAAC CTGGTGCGACGCCTTCT GCAGCATTCGTGGCAAA CGCGTGGATCTGGGTTG CGCGGCCACCTGTCCGA CGGTTAAAACCGGCGTG GATATTCAATGTTGCTCG ACCGATAACTGTAACCCG TTCCCGACCCGTAAACG CCCG	213	56.8	0.89	3099	290
B	QVNCLLPPKA GPCKGSFARY YFDIESGSCK AFIYGGCQPN SNNFAKRHHC EKRCRRFALG I	61	CAGGTTAACTGCCTGCT GCCGCCGAAAGCGGGCC CATGCAAAGGCAGCTTT GCCCGTTATTATTTTGAT ATTGAGAGTGGTAGCTGT AAAGCATTCATTTATGGT GGCTGTCAACCGAACAG CAACAACTTCGCAAAACG CCATCACTGCGAAAAAC GCTGCCGCCGTTTTGCG TTGGGCATT	183	50.3	0.90	-	260
C	HPTKPCMYCS FGQCVGPHIC CGPTGCEMG TAEANMCSEE DEDPIPCQVF GSDCALNNPD NIHGHCVADGI CCVDDTCTTH LGCL	84	CATCCGACGAAACCGTG TATGTACTGTAGCTTTGG TCAATGTGTTGGTCCGCA TATCTGCTGTGGTCCGAC GGGCTGTGAAATGGGCA CCGCGGAAGCCAACATG TGCAGCGAAGAAGATGA AGACCCGATTCCGTGCC AGGTCTTTGGCTCTGATT GTGCACTGAACAATCCG GACAATATTCATGGCCAC TGTGTGGCGGATGGTAT CTGCTGTGTTGATGACAC GTGCACCACGCATCTGG GTTGTCTG	252	53.2	0.90	-	329

Table S3. 2| Primer sequences used to assembly of gene A using different approaches for gene synthesis.

Gene synthesis strategy	Number of primers	Primer name	Primer sequence (5'-3')	Primer size (nt)
A1	2	Fw_A1	GGCAAACGCGTGGATCTGGGTTGCGCGGCCACCTGTCCGACGGTTAAAACC GGCGTGGATATTCAATGTTGCTCGACCGATAACTGTAACCCGTTCCCGACCC GTAAACGCCCGGGATCCCCGGGTACCGAGCTC	135
		Rv_A1	ATCCACGCGTTTGCCACGAATGCTGCAGAAGGCGTCGCACCAGGTTTTCGTG TAGCAGACGTGACCGTTCGGGCAATCTTTTGAGGTAATGTCCGGCGTGATGA AGCAGCGGATCCATATGACTAGTGGATCCTC	135
A2	14	F1_A2	ATCCACGCGTTTGCCACGAATGCTGCAGAAGGCGTCGC	38
		F2_A2	ACCAGGTTTTCGTGTAGCAGACGTGACCGTTCGGGCAATC	40
		F3_A2	TTTTGAGGTAATGTCCGGCGTGATGAAGCAGCGGATCCAT	40
		F4_A2	CCATATGACTAGTGGATCCT	20
		F5_A2	GGCAAACGCGTGGATCTGGGTTGCGCG	28
		F6_A2	GCCACCTGTCCGACGGTTAAAACCGGCGTGGATATTCAAT	40
		F7_A2	GTTGCTCGACCGATAACTGTAACCCGTTCCCGACCCGTAA	40
		F8_A2	ACCCGTAAACGCCCGGGATCCCCGGGTACCGAGCTC	36
		R1_A2	GAGGATCCACTAGTCATATGGATCCGCTGCTTCATCA	37
		R2_A2	CGCCGGACATTACCTCAAAAGATTGCCCGAACGGTCACGT	40
		R3_A2	CTGCTACACGAAAACCTGGTGCGACGCCTTCTGCAGCATT	40
		R4_A2	CGGGCGTTTACGGGTCGGGAACGGGTT	27
		R5_A2	ACAGTTATCGGTCGAGCAACATTGAATATCCACGCCGGTT	40
		R6_A2	TTAACCGTCGGACAGGTGGCCGCGCAACCCAGATCCACGC	40
A3	12	F1_A3	GCGTGATGAAGCAGCGGATCCATATGACTAGTGGATCCTC	40
		F2_A3	ATCTTTTGAGGTAATGTCCGGCGTGATGAAGCAGCGGATC	40
		F3_A3	CAGACGTGACCGTTCGGGCAATCTTTTGAGGTAATGTCCG	40
		F4_A3	CGCACCAGGTTTTCGTGTAGCAGACGTGACCGTTCGGGCA	40
		F5_A3	ACGAATGCTGCAGAAGGCGTCGCACCAGGTTTTCGTGTAG	40
		F6_A3	ATCCACGCGTTTGCCACGAATGCTGCAGAAGGCGT	35
		R1_A3	GGCAAACGCGTGGATCTGGGTTGCGCGGCCACCTGTCCGA	40
		R2_A3	TTGCGCGGCCACCTGTCCGACGGTTAAAACCGGCGTGGAT	40
		R3_A3	CGGTTAAAACCGGCGTGGATATTCAATGTTGCTCGACCGA	40
		R4_A3	ATTCAATGTTGCTCGACCGATAACTGTAACCCGTTCCCGA	40
		R5_A3	TAACTGTAACCCGTTCCCGACCCGTAAACGCCCGGGATCC	40
		R6_A3	CCCGTAAACGCCCGGGATCCCCGGGTACCGAGCTC	35
B1	14	F1_B1	ACAAGTTTGTACAAAAAAGCAGGCTTAGAAAACCT	35
		F2_B1	GTACTTCCAGATCCGCTGCTTCATCACGCCGGACATTACC	40
		F3_B1	TCAAAAGATTGCCCGAACGGTCACGTCTGCTACACGAAAA	40
		F4_B1	CCTGGTGCGACGCCTTCTGCAGCATTCGTGGCAAACGCGT	40
		F5_B1	GGATCTGGGTTGCGCGGCCACCTGTCCGACGGTTAAAACC	40
		F6_B1	GGCGTGGATATTCAATGTTGCTCGACCGATAACTGTAACC	40
		F7_B1	CGTTCCCGACCCGTAAACGCCCGTAATAAGACCCAGCTTT	40
		R1_B1	ACCACTTTGTACAAGAAAGCTGGGTCTTATTACGG	35
		R2_B1	GCGTTTACGGGTCGGGAACGGGTTACAGTTATCGGTCGAG	40
		R3_B1	CAACATTGAATATCCACGCCGGTTTTAACCGTCGGACAGG	40
		R4_B1	TGGCCGCGCAACCCAGATCCACGCGTTTGCCACGAATGCT	40
		R5_B1	GCAGAAGGCGTCGCACCAGGTTTTCGTGTAGCAGACGTGA	40
		R6_B1	CCGTTCGGGCAATCTTTTGAGGTAATGTCCGGCGTGATGA	40
		R7_B1	AGCAGCGGATCTGGAAGTACAGGTTTTCTAAGCCTGCTTT	40
B2	12	F1_B2	ACAAGTTTGTACAAAAAAGCAGGCTTAGAA	30
		F2_B2	TCCAGATCCGCTGCTTCATCACGCCGGACATTACCTCAAA	40
		F3_B2	AACGGTCACGTCTGCTACACGAAAACCTGGTGCGACGCCT	40
		F4_B2	TCGTGGCAAACGCGTGGATCTGGGTTGCGCGGCCACCTGT	40
		F5_B2	AAACCGGCGTGGATATTCAATGTTGCTCGACCGATAACTG	40
		F6_B2	CCGACCCGTAAACGCCCGTAATAAGACCCAGCTTTCTTGT	40
		R1_B2	ACCACTTTGTACAAGAAAGCTGGGTCTTAT	30
		R2_B2	TACGGGTCGGGAACGGGTTACAGTTATCGGTCGAGCAACA	40
		R3_B2	ACGCCGGTTTTAACCGTCGGACAGGTGGCCGCGCAACCCA	40
		R4_B2	TTTGCCACGAATGCTGCAGAAGGCGTCGCACCAGGTTTTC	40

		R5_B2	CGTGACCGTTCGGGCAATCTTTTGAGGTAATGTCCGGCGT	40
		R6_B2	CGGATCTGGAAGTACAGGTTTTCTAAGCCTGCTTTTTTGT	40
B3	8	F1_B2	ACAAGTTTGTACAAAAAAGCAGGCTTAGAAAACCT	35
		F2_B2	TCATCACGCCGGACATTACCTCAAAAGATTGCCCGAACGGTCACGTCTGCTACACGAAAA	60
		F3_B2	AGCATTCGTGGCAAACGCGTGGATCTGGGTTGCGCGGCCACCTGTCCGACGGTTAAAACC	60
		F4_B2	CTCGACCGATAACTGTAACCCGTTCCCGACCCGTAAACGCCCGTAATAAGACCCAGCTTT	60
		R1_B2	ACCACTTTGTACAAGAAAGCTGGGTCTTATTACGG	35
		R2_B2	GGTTACAGTTATCGGTCGAGCAACATTGAATATCCACGCCGGTTTTAACCGTCGGACAGG	60
		R3_B2	ACGCGTTTGCCACGAATGCTGCAGAAGGCGTCGCACCAGGTTTTCGTGTAGCAGACGTGA	60
		R4_B2	GGTAATGTCCGGCGTGATGAAGCAGCGGATCTGGAAGTACAGGTTTTCTAAGCCTGCTTT	60

Table S3. 3| Primer sequences used to assembly of genes B and C.

Gene name	Gene assembly strategy	Number of primers	Primer name	Primer sequence (5'-3')	Primer size (nt)
B	B3	6	B_F1	ACAAGTTTGTACAAAAAAGCAGGCTTAGAAAACCTGTACTTCCAGCAGGTTAACTGCCTG	60
			B_F2	ATGCAAAGGCAGCTTTGCCCGTTATTATTTTGATATTGAGAGTGGTAGCTGTAAAGCATT	60
			B_F3	CGAACAGCAACAACTTCGCAAAACGCCATCACTGCGAAAAACGCTGCCGCCGTTTTGCGT	60
			B_R1	ACCACTTTGTACAAGAAAGCTGGGTCTTATTAAATGCCCAACGCAAAACGGCGGCAGCGT	60
			B_R2	TGCGAAGTTGTTGCTGTTCGGTTGACAGCCACCATAAATGAATGCTTTACAGCTACCACT	60
			B_R3	GGGCAAAGCTGCCTTTGCATGGGCCCGCTTTCGGCGGCAGCAGGCAGTTAACCTGCTGGA	60
C	B3	8	C_F1	GGGGACAAGTTTGTACAAAAAAGCAGGCTTAGAAAACCTGTACTTCCAGCATCCGACGA	59
			C_F2	TTTGGTCAATGTGTTGGTCCGCATATCTGCTGTGGTCCGACGGGCTGTGAAATGGGCACC	60
			C_F3	CGAAGAAGATGAAGACCCGATTCCGTGCCAGGTCTTTGGCTCTGATTGTGCACTGAACAA	60
			C_F4	ACTGTGTGGCGGATGGTATCTGCTGTGTTGATGACACGTGCACCACGCATCTGGGTTGTC	60
			C_R1	CCCCACCACTTTGTACAAGAAAGCTGGGTCTTATTACAGACAACCCAGATGCGTGGTG	58
			C_R2	GATACCATCCGCCACACAGTGGCCATGAATATTGTCCGGATTGTTCAGTGCACAATCAGA	60
			C_R3	TCGGGTCTTCATCTTCTTCGCTGCACATGTTGGCTTCCGCGGTGCCCATTTCACAGCCCG	60
			C_R4	GGACCAACACATTGACCAAAGCTACAGTACATACACGGTTTCGTCGGATGCTGGAAGTAC	60

Table S3. 4| PCR programs used to optimize the PCR profile of a PCR-assembly reaction.

PCR program	Number of cycles	Temperature	Time
I	1	95°C	2 m
	22/24/26	95°C	20 s
		60°C	10 s
		70°C	1 s
II	1	95°C	2 m
	22/24/26	95°C	16 s
		60°C	10 s
		70°C	1 s
III	1	95°C	2 m
	22/24/26	95°C	20 s
		60°C	8 s
		70°C	1 s
IV	1	95°C	2 m
	22/24/26	95°C	16 s
		60°C	8 s
		70°C	1 s
V	1	95°C	2 m
	22/24/26	95°C	20 s
		60°C	10 s
		70°C	3 s
VI	1	95°C	2 m
	22/24/26	95°C	16 s
		60°C	10 s
		70°C	3 s
VII	1	95°C	2 m
	22/24/26	95°C	20 s
		60°C	8 s
		70°C	3 s
VIII	1	95°C	2 m
	22/24/26	95°C	16 s
		60°C	8 s
		70°C	3 s

Table S3. 5| Sequence properties of 96 genes synthesised using the HTP gene synthesis platform.

Gene	Protein size (aa)	Gene size (nt)	Recombinant gene size (nt)	GC content (%)	CAI
1	60	180	233	47.2	0.90
2	67	201	254	47.8	0.88
3	67	201	254	51.2	0.89
4	60	180	233	51.1	0.89
5	60	180	233	47.2	0.88
6	60	180	233	45.0	0.85
7	60	180	233	47.2	0.85
8	64	192	245	50.2	0.82
9	64	192	245	51.4	0.82
10	65	195	248	50.4	0.91
11	66	198	251	48.2	0.86
12	66	198	251	47.0	0.84
13	66	198	251	48.6	0.85
14	58	174	227	51.7	0.84
15	58	174	227	53.3	0.86
16	63	189	242	40.2	0.84
17	64	192	245	41.2	0.90
18	54	162	215	48.8	0.84
19	60	180	233	48.1	0.92
20	59	177	230	51.3	0.90
21	65	195	248	50.8	0.97
22	62	186	239	46.0	0.87
23	62	186	239	48.9	0.86
24	63	189	242	45.4	0.86
25	61	183	236	47.5	0.85
26	62	186	239	46.4	0.85
27	66	198	251	47.0	0.88
28	66	198	251	48.2	0.88
29	66	198	251	44.5	0.86
30	66	198	251	46.6	0.87
31	65	195	248	48.0	0.91
32	65	195	248	44.6	0.89
33	65	195	248	49.6	0.89
34	63	189	242	47.6	0.82
35	55	165	218	53.7	0.90
36	63	189	242	49.6	0.83
37	65	195	248	56.5	0.89
38	62	186	239	52.7	0.85
39	55	165	218	56.4	0.90
40	62	186	239	55.2	0.90
41	62	186	239	44.6	0.90
42	62	186	239	46.8	0.90
43	54	162	215	52.0	0.93
44	64	192	245	48.4	0.83
45	61	183	236	47.0	0.85
46	61	183	236	47.5	0.85
47	67	201	254	50.0	0.89
48	67	201	254	43.8	0.86
49	67	201	254	46.8	0.86
50	60	180	233	51.1	0.86
51	60	180	233	50.0	0.80
52	60	180	233	51.5	0.81
53	60	180	233	47.8	0.88
54	60	180	233	47.6	0.90
55	66	198	251	44.2	0.85
56	67	201	254	44.2	0.84
57	61	183	236	48.3	0.84

58	61	183	236	48.1	0.84
59	56	168	221	52.5	0.92
60	62	186	239	47.3	0.83
61	60	180	233	46.1	0.83
62	60	180	233	46.8	0.84
63	63	189	242	50.0	0.83
64	64	192	245	48.4	0.84
65	61	183	236	51.7	0.83
66	61	183	236	50.8	0.83
67	62	186	239	52.3	0.92
68	57	171	224	45.6	0.84
69	57	171	224	48.2	0.85
70	67	201	254	50.3	0.82
71	63	189	242	56.2	0.93
72	62	186	239	54.3	0.84
73	62	186	239	54.0	0.84
74	61	183	236	51.3	0.93
75	62	186	239	44.0	0.88
76	64	192	245	55.2	0.90
77	63	189	242	57.0	0.93
78	66	198	251	48.2	0.87
79	55	165	218	49.0	0.94
80	65	195	248	52.3	0.83
81	64	192	245	53.5	0.88
82	61	183	236	45.4	0.85
83	61	183	236	46.6	0.86
84	65	195	248	50.0	0.94
85	66	198	251	48.5	0.85
86	60	180	233	42.2	0.82
87	58	174	227	43.1	0.89
88	67	201	254	50.8	0.84
89	65	195	248	51.8	0.82
90	62	186	239	43.0	0.88
91	65	195	248	46.7	0.85
92	67	201	254	49.8	0.89
93	65	195	248	50.8	0.84
94	65	195	248	51.3	0.90
95	66	198	251	52.5	0.89
96	60	180	233	48.3	0.82

Supplemental information – Chapter 4

Table S4. 1| Sequence properties of seven endonucleases produced in *E. coli*.

Mismatch cleavage endonuclease	Origin	Organism	Gene sequence	Gene size (nt)	Protein primary sequence	Protein size (aa)	Molecular weight (kDa)	Recombinant protein primary sequence	Recomb. protein size (aa)	Recomb. molecular weight (kDa)
Endonuclease V	Bacteria	*Escherichia coli*	GATCTCGCGTCATTACGCGCTCAACAAAT CGAACTGGCTTCTTCTGTGATCCGCGAGG ATCGACTCGATAAAGATCCACCGGATCTG ATCGCCGGAGCCGATGTCGGGTTTGAGC AGGGCGGAGAAGTGACGCGAGCGGCGAT GGTGCTGCTGAAATATCCCTCGCTTGAGC TGGTCGAGTATAAAGTTGCCCGCATCGCC ACCACCATGCCTTACATTCCAGGTTTTCTT TCCTTCCGCGAATATCCTGCGCTGCTGGC AGCGTGGGAGATGCTGTCGCAAAAGCCG GATTTAGTGTTTGTCGATGGTCATGGGAT CTCGCATCCTCGCCGTCTTGGCGTCGCCA GCCATTTTGGCTTATTGGTGGATGTGCCG ACCATTGGCGTGGCGAAAAAACGGCTCTG CGGTAAATTCGAACCGCTCTCCAGCGAAC CGGGCGCGCTGGCCCCACTGATGGATAA AGGCGAGCAGCTGGCCTGGGTCTGGCGC AGCAAAGCGCGCTGTAACCCGTTGTTTAT CGCTACCGGCCATCGGGTCAGCGTGGAC AGCGCGCTGGCGTGGGTACAACGCTGCA TGAAAGGCTATCGTCTGCCGGAGCCAaCG CGCTGGGCGGACGCGGTGGCCTCGGAAC GTCCGGCGTTCGTGCGCTATACAGCAAAT CAGCCC	666	DLASLRAQQIELASSVIREDRLD KDPPDLIAGADVGFEQGGEVTR AAMVLLKYPSLELVEYKVARIAT TMPYIPGFLSFREYPALLAAWE MLSQKPDLVFVDGHGISHPRRL GVASHFGLLVDVPTIGVAKKRL CGKFEPLSSEPGALAPLMDKG EQLAWVWRSKARCNPLFIATG HRVSVDSALAWVQRCMKGYRL PEPTRWADAVASERPAFVRYT ANQP	222	24.54	MGSSHHHHHHSSGPQQGLRD LASLRAQQIELASSVIREDRLDK DPPDLIAGADVGFEQGGEVTRA AMVLLKYPSLELVEYKVARIATT MPYIPGFLSFREYPALLAAWEM LSQKPDLVFVDGHGISHPRRLG VASHFGLLVDVPTIGVAKKRLC GKFEPLSSEPGALAPLMDKGE QLAWVWRSKARCNPLFIATGH RVSVDSALAWVQRCMKGYRLP EPTRWADAVASERPAFVRYTA NQP	241	26.64
Endonuclease III-wt	Eubateria	Eubacterium SCB49	TGGGGCAAGAATGGCCACCGTACCGTTG GCGCCATTGCCGAGGCACACCTGTCGAA AAAAGCCCAGAAAGCAATTGATAAACTGC TGAATGGCAAGAGCCTGGCACTCGTTAGC ACCTTTGGCGATGAAATCCGTTCGGACAA AAAGTATCGCTCCTTTGCGCCATGGCATT ATGTGAGTTTCCCATTTGAAGCCACCTAC GATACCCACCCAAAATCCGAGAAAGGCGA CGTGATTACCGGTATCAACACCTGCATTG AAAAAATCAAAGATGAGAATAGTACTCGTG AAGACAAAGCATTCTATCTGAAAATGTTAG TCCATTTCATCGGCGATATCCACCAACCA CTGCACGTCGGCCTGGCAGAAGACAAGG GCGGTAACACCTTCCAGGTGCAATGGTTC GATCAGGGCACCAACCTGCACAGCGTCT GGGATACCAAAATTATTGAATCCTACGAG ATGTCCTACACTGAATTAGCGGATAACCG CAAACGCCTGACCAAAGCCGAAATTGCCA CCATCCAGCTGGGCGACGCAAAAACCTG GGCCGCAGAAAGCCGCGAACTCTGTAAA GATATTTACGCAAACACGAAACCGGGCGA AAACCTGGGTTACCGCTACATGTTCGATT ATATGGATGTCACCCGCACCCAACTGCAA AAAGGCGGCATCCGCCTGGCAACCGTGC TGAACGAAATTTTTGGC	708	WGKNGHRTVGAIAEAHLSKKA QKAIDKLLNGKSLALVSTFGDEI RSDKKYRSFAPWHYVSFPFEA TYDTHPKSEKGDVITGINTCIEKI KDENSTREDKAFYLKMLVHFIG DIHQPLHVGLAEDKGGNTFQV QWFDQGTNLHSVWDTKIIESYE MSYTELADNRKRLTKAEIATIQL GDAKTWAAESRELCKDIYANTK PGENLGYRYMFDYMDVTRTQL QKGGIRLATVLNEIFG	236	26.83	MGSSHHHHHHSSGPQQGLRW GKNGHRTVGAIAEAHLSKKAQK AIDKLLNGKSLALVSTFGDEIRS DKKYRSFAPWHYVSFPFEATY DTHPKSEKGDVITGINTCIEKIKD ENSTREDKAFYLKMLVHFIGDIH QPLHVGLAEDKGGNTFQVQWF DQGTNLHSVWDTKIIESYEMSY TELADNRKRLTKAEIATIQLGDA KTWAAESRELCKDIYANTKPGE NLGYRYMFDYMDVTRTQLQKG GIRLATVLNEIFG	255	28.93
Endonuclease III-mut	Eubacteria	Eubacterium SCB49	TGGGGCAAGAATGGCCACCGTACCGTTG GCGCCATTGCCGAGGCACACCTGTCGAA AAAAGCCCAGAAAGCAATTGATAAACTGC TGAATGGCAAGAGCCTGGCACTCGTTAGC ACCTTTGGCGATGAAATCCGTTCGGACAA AAAGTATCGCTCCTTTGCGCCATGGCATT ATGTGAGTTTCCCATTTGAAGCCACCTAC GATACCCACCCAAAATCCGAGAAAGGCGA CGTGATTACCGGTATCAACACCTGCATTG AAAAAATCAAAGATGAGAATAGTACTCGTG AAGACAAAGCATTCTATCTGAAAATGTTAG TCCATTTCATCGGCGATATCCACCAACCA CTGCACGTCGGCTTCCTGGGCGATGAAG GCGGTAACACCTTCCAGGTGCAATGGTTC GATCAGGGCACCAACCTGCACAGCGTCT	708	WGKNGHRTVGAIAEAHLSKKA QKAIDKLLNGKSLALVSTFGDEI RSDKKYRSFAPWHYVSFPFEA TYDTHPKSEKGDVITGINTCIEKI KDENSTREDKAFYLKMLVHFIG DIHQPLHVGFLGDEGGNTFQV QWFDQGTNLHSVWDTKIIESYE MSYTELADNRKRLTKAEIATIQL GDAKTWAAESRELCKDIYANTK PGENLGYRYMFDYMDVTRTQL QKGGIRLATVLNEIFG	236	26.84	MGSSHHHHHHSSGPQQGLRW GKNGHRTVGAIAEAHLSKKAQK AIDKLLNGKSLALVSTFGDEIRS DKKYRSFAPWHYVSFPFEATY DTHPKSEKGDVITGINTCIEKIKD ENSTREDKAFYLKMLVHFIGDIH QPLHVGFLGDEGGNTFQVQWF DQGTNLHSVWDTKIIESYEMSY TELADNRKRLTKAEIATIQLGDA KTWAAESRELCKDIYANTKPGE NLGYRYMFDYMDVTRTQLQKG GIRLATVLNEIFG	255	28.92

			GGGATACCAAAATTATTGAATCCTACGAG ATGTCCTACACTGAATTAGCGGATAACCG CAAACGCCTGACCAAAGCCGAAATTGCCA CCATCCAGCTGGGCGACGCAAAAACCTG GGCCGCAGAAAGCCGCGAACTCTGTAAA GATATTTACGCAAACACGAAACCGGGCGA AAACCTGGGTTACCGCTACATGTTCGATT ATATGGATGTCACCCGCACCCAACTGCAA AAAGGCGGCATCCGCCTGGCAACCGTGC TGAACGAAATTTTTGGC							
Endonuclease I	Plant	*Apium graveolens*	TGGGGCAAACAGGGTCATTTTGCGATTTG CAAAATCGCCCAAGGCTTCCTGTCTAAAG ATGCACTGACCGCTGTCAAAGCGCTGCTG CCGGAATACGCGGATGGTGACCTGGCGG CCGTTTGTAGCTGGGCCGACGAAGTCCGT TTTCATATGCGCTGGAGCTCTCCGCTGCA CTATGTTGATACGCCGGACTTCCGTTGCA ATTATAAATACTGCCGCGATTGTCATGACA GCGTGGGCCGTAAAGATCGCTGTGTTACC GGTGCAATTCATAATTACACGGAACAGCT GCTGCTGGGCGTGCACGACCTGAACTCG AAAATGAACAATAACCTGACCGAAGCTCT GATGTTTCTGAGCCACTTCGTCGGTGATG TGCATCAACCGCTGCACGTTGGCTTTCTG GGTGACGAAGGCGGTAATACCATTACGGT TCGTTGGTACCGTCGCAAAACCAACCTGC ATCACGTCTGGGATACGATGATGATCGAA AGTTCCCTGAAAACCTTCTATAATTCCGAC CTGTCATCGCTGATTCAGGCAATCCAATC AAACATTACCGGCGTGTGGCTGACGGATA GTCTGTCCTGGTCAAATTGCACCGCAGAT CATGTGGTTTGTCCGGACCCGTATGCTTC GGAAAGCATCGAACTGGCCTGCAAATTTG CATATCGTAATGCTACGCCGGGTACCACC CTGGGTGATGAATATTTCCTGAGTCGTCT GCCGGTTGCGGAAAAACGCCTGGCCCAG GCCGGTGTGCGTCTGGCAGCTACCCTGA ATCGCATCTTTACGTCTAACCCGAGTGAT CTGACCCGTCTGAATATGCATAACGGCGG TCACCGCAGCTCTAATAACATTGAAATCGT T	870	WGKQGHFAICKIAQGFLSKDAL TAVKALLPEYADGDLAAVCSWA DEVRFHMRWSSPLHYVDTPDF RCNYKYCRDCHDSVGRKDRCV TGAIHNYTEQLLLGVHDLNSKM NNNLTEALMFLSHFVGDVHQPL HVGFLGDEGGNTITVRWYRRK TNLHHVWDTMMIESSLKTFYNS DLSSLIQAIQSNITGVWLTDSLS WSNCTADHVVCPDPYASESIEL ACKFAYRNATPGTTLGDEYFLS RLPVAEKRLAQAGVRLAATLNR IFTSNPSDLTRLNMHNGGHRSS NNIEIV	290	32.60	MGSSHHHHHHSSGPQQGLRW GKQGHFAICKIAQGFLSKDALT AVKALLPEYADGDLAAVCSWA DEVRFHMRWSSPLHYVDTPDF RCNYKYCRDCHDSVGRKDRCV TGAIHNYTEQLLLGVHDLNSKM NNNLTEALMFLSHFVGDVHQPL HVGFLGDEGGNTITVRWYRRK TNLHHVWDTMMIESSLKTFYNS DLSSLIQAIQSNITGVWLTDSLS WSNCTADHVVCPDPYASESIEL ACKFAYRNATPGTTLGDEYFLS RLPVAEKRLAQAGVRLAATLNR IFTSNPSDLTRLNMHNGGHRSS NNIEIV	309	34.70
Endonuclease II	Fungi	*Tulasnella calospora*	TGGGGCGGTCTGGGTCATAAAACCGTCG CGAACGTGGCCCTGCAATTTCTGCAGGAA GATGCCCTGGCCGGTGTTGAAGCGGTCC TGGCCGCCGATGGCCACAAAGAATCAGA CAATCCGTCGATTGTGGATGTTGCGACCT GGGCAGACGCTTTTGGTCGTAAGAAAGGC GGCACGTTCAGCAAAAAATTCCATTACATC AACGCCCACGATGACCCGCCGTTCGAATG CAATGTGGATCTGGAACGTGACTGCAGTG AAAAAGGCGAATGTATTGTTAAAGCGATC GCCAACTACACCCAACGTCTGATTCGCCC GTCCCGTAACGTTAATGATACGGCAGACG CTCTGAAATTTCTGGTCCATTTCATTGGTG ATATCACCCAGCCGCTGCACACGGAAGAC AAAGAACGCGGCGGTAATGGCATTCCGGT CGCGTGGAACGGCAATGGTAACAAAAATC TGCATAGCGTCTGGGATACCAACATGGTG AAAAAACTGGCTGGTTCTGATAATGAAGA AAACCTGAATGCATGGACCGACATTATCG TTAACGAAATTAACAATGGCAGCTATAAAC CGCTGATCCCGGAATGGCTGAACTGCACC GATCCGCGTGACGCACTGAATTGTGCTCT GAAATGGGCGACGGATAGCAACTCTTTTA TTTGTACCTACGTGCTGAAAAATGATACG GACGGCTGGGAACTGAGTGGTTCCTATTA CCGCGGCGCAGCTCCGATTATCCAGCAA CAGATCGCCAAAGGCGGTGTGCGTCTGG CAGTTTGGCTGAACCAGCTGTTCGGCTCT GGTGAACCGGGCAAACTGCCGCTGGATC GCGCGCAAGTTATTATCCAGAAT	861	WGGLGHKTVANVALQFLQEDA LAGVEAVLAADGHKESDNPSIV DVATWADAFGRKKGGTFSKKF HYINAHDDPPFECNVDLERDCS EKGECIVKAIANYTQRLIRPSRN VNDTADALKFLVHFIGDITQPLH TEDKERGGNGIPVAWNGNGNK NLHSVWDTNMVKKLAGSDNEE NLNAWTDIIVNEINNGSYKPLIP EWLNCTDPRDALNCALKWATD SNSFICTYVLKNDTDGWELSGS YYRGAAPIIQQQIAKGGVRLAV WLNQLFGSGEPGKLPLDRAQVI IQN	287	31.60	MGSSHHHHHHSSGPQQGLRW GGLGHKTVANVALQFLQEDALA GVEAVLAADGHKESDNPSIVDV ATWADAFGRKKGGTFSKKFHYI NAHDDPPFECNVDLERDCSEK GECIVKAIANYTQRLIRPSRNVN DTADALKFLVHFIGDITQPLHTE DKERGGNGIPVAWNGNGNKNL HSVWDTNMVKKLAGSDNEENL NAWTDIIVNEINNGSYKPLIPEW LNCTDPRDALNCALKWATDSN SFICTYVLKNDTDGWELSGSYY RGAAPIIQQQIAKGGVRLAVWL NQLFGSGEPGKLPLDRAQVIIQ N	306	33.70
T7 Endonuclease I-6HIS	Enterobacteria	Bacteriophage T7	GCGGGTTATGGCGCCAAGGGTATCCGTA AAGTCGGCGCGTTCCGCTCCGGCCTGGA AGATAAGGTCAGCAAACAGCTGGAGAGCA AAGGCATCAAGTTTGAATACGAAGAGTGG AAAGTCCCGTATGTTATCCCGGCAAGCAA TCATACCTACACCCCGGATTTCCTGCTGC CGAATGGTATCTTCGTTGAGACCAAAGGC	444	AGYGAKGIRKVGAFRSGLEDKV SKQLESKGIKFEYEEWKVPYVI PASNHTYTPDFLLPNGIFVETK GLWESDDRKKHLLIREQHPELD IRIVFSSSRTKLYKGSPTSYGEF CEKHGIKFADKLIPAEWIKEPKK EVPFDRLKRKGGKK	148	17.04	MGSSHHHHHHSSGPQQGLRA GYGAKGIRKVGAFRSGLEDKVS KQLESKGIKFEYEEWKVPYVIP ASNHTYTPDFLLPNGIFVETKGL WESDDRKKHLLIREQHPELDIRI VFSSSRTKLYKGSPTSYGEFCE	167	19.14

			CTGTGGGAGAGCGATGATCGTAAAAAACA TCTGTTGATCCGTGAACAACACCCGGAGC TGGATATTCGCATCGTGTTCAGCAGTAGC CGTACTAAACTTTATAAGGGTAGTCCGAC GAGCTATGGCGAGTTCTGTGAAAAACACG GCATTAAGTTTGCCGATAAGTTAATCCCG GCCGAATGGATTAAAGAACCGAAAAAAGA AGTTCCGTTTGATCGCTTGAAACGCAAAG GTGGCAAGAAA					KHGIKFADKLIPAEWIKEPKKEV PFDRLKRKGGKK		
T7 Endonuclease I-MBP	Enterobacteria	Bacteriophage T7	GCGGGTTATGGCGCCAAGGGTATCCGTA AAGTCGGCGCGTTCCGCTCCGGCCTGGA AGATAAGGTCAGCAAACAGCTGGAGAGCA AAGGCATCAAGTTTGAATACGAAGAGTGG AAAGTCCCGTATGTTATCCCGGCAAGCAA TCATACCTACACCCCGGATTTCCTGCTGC CGAATGGTATCTTCGTTGAGACCAAAGGC CTGTGGGAGAGCGATGATCGTAAAAAACA TCTGTTGATCCGTGAACAACACCCGGAGC TGGATATTCGCATCGTGTTCAGCAGTAGC CGTACTAAACTTTATAAGGGTAGTCCGAC GAGCTATGGCGAGTTCTGTGAAAAACACG GCATTAAGTTTGCCGATAAGTTAATCCCG GCCGAATGGATTAAAGAACCGAAAAAAGA AGTTCCGTTTGATCGCTTGAAACGCAAAG GTGGCAAGAAA	444	AGYGAKGIRKVGAFRSGLEDKV SKQLESKGIKFEYEEWKVPYVI PASNHTYTPDFLLPNGIFVETK GLWESDDRKKHLLIREQHPELD IRIVFSSSRTKLYKGSPTSYGEF CEKHGIKFADKLIPAEWIKEPKK EVPFDRLKRKGGKK	148	17.04	MGKKGFMLFTLLAAFSGFAQAK IEEGKLVIWINGDKGYNGLAEV GKKFEKDTGIKVTVEHPDKLEE KFPQVAATGDGPDIIFWAHDRF GGYAQSGLLAEITPDKAFQDKL YPFTWDAVRYNGKLIAYPIAVE ALSLIYNKDLLPNPPKTWEEIPA LDKELKAKGKSALMFNLQEPYF TWPLIAADGGYAFKYENGKYDI KDVGVDNAGAKAGLTFLVDLIK NKHMNADTDYSIAEAAFNKGET AMTINGPWAWSNIDTSKVNYG VTVLPTFKGQPSKPFVGVLSAG INAASPNKELAKEFLENYLLTDE GLEAVNKDKPLGAVALKSYEEE LAKDPRIAATMENAQKGEIMPNI PQMSAFWYAVRTAVINAASGR QTVDEALKDAQTSMGSSHHHH HHSSGPQQGLRAGYGAKGIRK VGAFRSGLEDKVSKQLESKGIK FEYEEWKVPYVIPASNHTYTPD FLLPNGIFVETKGLWESDDRKK HLLIREQHPELDIRIVFSSSRTKL YKGSPTSYGEFCEKHGIKFADK LIPAEWIKEPKKEVPFDRLKRKG GKK	555	61.64

Table S4. 2| Sequence properties of *lacZ-gfp* gene synthesised using overlapping oligonucleotides.

Gene construct	Gene sequence	Gene size (nt)	GC content (%)	CAI
lacZ plus GFP	TCAGCAAGGGCTGAGGGCGCCCAATACGCAAACCGCCTCTCCCCGCGCGT TGGCCGATTCATTAATGCAGCTGGCACGACAGGTTTCCCGACTGGAAAGCG GGCAGTGAGCGCAACGCAATTAATGTGAGTTAGCTCACTCATTAGGCACCC CAGGCTTTACACTTTATGCTTCCGGCTCGTATGTTGTGTGGAATTGTGAGCG GATAACAATTTCACACAGGAAACAGCTATGGTTAGCAAAGGTGAAGAACTGT TTACGGGCGTTGTGCCGATCCTGGTGGAACTGGACGGTGATGTTAATGGTC ATAAATTCTCTGTGAGTGGCGAAGGTGAAGGCGATGCGACCTATGGTAAAC TGACGCTGAAATTTATTTGCACCACCGGTAAACTGCCGGTTCCGTGGCCGA CCCTGGTCACCACCCTGACCTACGGTGTGCAGTGTTTCGCACGCTATCCGG ATCATATGAAACAACACGACTTTTTCAAAAGCGCTATGCCGGAAGGTTACGT TCAGGAACGTACCATTTTCTTTAAAGATGACGGCAACTACAAAACCCGCGCC GAAGTCAAATTTGAAGGTGATACGCTGGTGAACCGTATTGAACTGAAAGGC ATCGATTTCAAAGAAGACGGTAATATCCTGGGCCATAAACTGGAATACAACT ACAACTCACACAAAGTTTACATTACCGCGGATAAACAGAAAAACGGTATCAA AGTCAACTTCAAAACGCGTCATAACATCGAAGATGGCTCTGTGCAACTGGC CGACCACTACCAGCAAAACACCCCGATCGGTGATGGCCCGGTTCTGCTGC CGGACAATCATTATCTGTCCACCCAGTCAGCACTGTCGAAAGATCCGAATG AAAAACGCGACCACATGGTGCTGCTGGAATTTGTTACCGCGGCCGGTATTA CGCTGGGCATGGATGAACTGTACAAATAATCAGCTTCCGCTGAGG	967	49	0.90

Table S4. 3| Primer sequences used to assembly of the *lacZ-gfp* gene.

Primer name	Primer sequence (5'-3')	Primer size (nt)
GFP_F1	TCAGCAAGGGCTGAGGGCGCCCAATACGCAAACCGCCTCTCCCCGCGCGTTGGCCGATTC	60
GFP_F2	AGGTTTCCCGACTGGAAAGCGGGCAGTGAGCGCAACGCAATTAATGTGAGTTAGCTCACT	60
GFP_F3	ACACTTTATGCTTCCGGCTCGTATGTTGTGTGGAATTGTGAGCGGATAACAATTTCACAC	60
GFP_F4	AAAGGTGAAGAACTGTTTACGGGCGTTGTGCCGATCCTGGTGGAACTGGACGGTGATGTT	60
GFP_F5	GAGTGGCGAAGGTGAAGGCGATGCGACCTATGGTAAACTGACGCTGAAATTTATTTGCAC	66
GFP_F6	CGTGGCCGACCCTGGTCACCACCCTGACCTACGGTGTGCAGTGTTTCGCACGCTATCCGG	60
GFP_F7	TTTTTCAAAAGCGCTATGCCGGAAGGTTACGTTCAGGAACGTACCATTTTCTTTAAAGAT	60
GFP_F8	CGCCGAAGTCAAATTTGAAGGTGATACGCTGGTGAACCGTATTGAACTGAAAGGCATCGA	60
GFP_F9	TCCTGGGCCATAAACTGGAATACAACTACAACTCACACAAAGTTTACATTACCGCGGATA	60
GFP_F10	GTCAACTTCAAAACGCGTCATAACATCGAAGATGGCTCTGTGCAACTGGCCGACCACTAC	60
GFP_F11	TGATGGCCCGGTTCTGCTGCCGGACAATCATTATCTGTCCACCCAGTCAGCACTGTCGAA	60
GFP_F12	ACCACATGGTGCTGCTGGAATTTGTTACCGCGGCCGGTATTACGCTGGGCATGGATGAAC	60
GFP_R1	TCAGCGGAAGCTGAGGTTATTTGTACAGTTCATCCATGCCCAGCGTAATACCGGCCGCGG	60
GFP_R2	TTCCAGCAGCACCATGTGGTCGCGTTTTTCATTCGGATCTTTCGACAGTGCTGACTGGGT	60
GFP_R3	GCAGCAGAACCGGGCCATCACCGATCGGGGTGTTTTGCTGGTAGTGGTCGGCCAGTTGCA	60
GFP_R4	TGACGCGTTTTGAAGTTGACTTTGATACCGTTTTTCTGTTTATCCGCGGTAATGTAAACT	60
GFP_R5	TTCCAGTTTATGGCCCAGGATATTACCGTCTTCTTTGAAATCGATGCCTTTCAGTTCAAT	60
GFP_R6	CTTCAAATTTGACTTCGGCGCGGGTTTTGTAGTTGCCGTCATCTTTAAAGAAAATGGTAC	60
GFP_R7	GGCATAGCGCTTTTGAAAAAGTCGTGTTGTTTCATATGATCCGGATAGCGTGCGAAACAC	60
GFP_R8	GGTGACCAGGGTCGGCCACGGAACCGGCAGTTTACCGGTGGTGCAAATAAATTTCAGCGT	60
GFP_R9	CGCCTTCACCTTCGCCACTCACAGAGAATTTATGACCATTAACATCACCGTCCAGTTCCA	60
GFP_R10	GTAAACAGTTCTTCACCTTTGCTAACCATAGCTGTTTCCTGTGTGAAATTGTTATCCGCT	60
GFP_R11	GAGCCGGAAGCATAAAGTGTAAAGCCTGGGGTGCCTAATGAGTGAGCTAACTCACATTAA	60
GFP_R12	GCTTTCCAGTCGGGAAACCTGTCGTGCCAGCTGCATTAATGAATCGGCCAACGCGCGGGG	60

Supplemental information – Chapter 5

Table S5. 1| Properties of venom peptides selected for this study and yield after recombinant expression of three gene variants per peptide.

Peptide	Primary sequence	Peptide name	Origin	Organism	Disulfide-bridges	Peptide size (aa)	Gene sequence, Variant A	GC content (%)	CAI	Gene sequence, Variant B	GC content (%)	CAI	Gene sequence, Variant C	GC content (%)	CAI	Gene size (nt)	Peptide concentration (mg/L culture)		
																	Variant A	Variant B	Variant C
S1	CTCKDMTDKECLYFCHQDIIW	Sarafotoxin-D	Snake	*Atractaspis engaddensis*	2	21	TGCACCTGTAAGGATATGACGGACAAAGAGTGCCTGTACTTCTGTCACCAGGACATTATCTGG	43.5	0.81	TGCACCTGTAAAGATATGACGGACAAGGAATGCCTGTATTTCTGCCATCAAGACATTATTTGG	37.7	0.85	TGCACCTGCAAAGATATGACCGATAAAGAGTGCTTATATTTTTGCCATCAGGATATCATTTGG	34.8	0.91	63	52.64	42.39	57.40
S2	CSCNDINDKECMYFCHQDVIWDEP	Sarafotoxin-m	Snake	*Atractaspis microlepidota*	2	24	TGCAGCTGCAACGACATTAATGATAAAGAGTGTATGTACTTCTGTCACCAGGATGTTATCTGGGACGAACCG	41.0	0.87	TGTAGCTGTAACGATATTAACGACAAGGAATGTATGTATTTCTGCCACCAGGATGTGATTTGGGACGAACCG	39.7	0.88	TGCAGTTGCAACGATATCAACGATAAAGAATGCATGTATTTTTGCCATCAAGACGTCATTTGGGATGAACCG	37.2	0.91	72	68.86	74.22	88.60
S3	KCLPPGKPCYGATQKIPCCGVCSHNKCT	Huwentoxin-X	Spider	*Ornithoctonus huwena*	3	28	AAATGTCTGCCACCGGGCAAACCGTGTTACGGTGCGACGCAGAAGATCCCGTGCTGCGGTGTTTGCAGCCACAACAAGTGTACC	53.3	0.84	AAATGTCTGCCGCCGGGCAAGCCGTGCTATGGTGCTACGCAAAAGATTCCGTGCTGTGGTGTCTGCTCGCATAACAAATGTACC	50.0	0.83	AAGTGTCTGCCACCGGGCAAACCGTGTTATGGCGCCACTCAGAAAATTCCGTGCTGTGGTGTGTGCTCACATAATAAATGCACC	47.1	0.87	84	75.81	80.19	84.54
S4	AEKDCIAPGAPCFGTDKPCCNPRAWCSSYANKCL	Ptu1	Insect	*Peirates turpis*	3	34	GCAGAGAAAGACTGCATCGCTCCGGGTGCGCCGTGTTTCGGCACCGATAAGCCGTGCTGTAATCCACGTGCGTGGTGTAGCAGCTACGCCAACAAATGCCTG	55.6	0.85	GCCGAAAAAGACTGTATTGCCCCGGGTGCTCCGTGCTTTGGTACCGACAAGCCGTGCTGTAATCCGCGTGCCTGGTGCTCGTCCTATGCGAACAAATGCCTG	54.6	0.87	GCCGAGAAAGACTGCATTGCGCCGGGCGCGCCATGCTTTGGCACTGATAAACCGTGCTGCAATCCACGTGCGTGGTGTAGCAGCTATGCGAATAAATGTCTG	52.8	0.88	102	105.28	66.25	99.27
S5	AVRIGPCDQVCPRIVPERHECCRAHGRSGY	Diapause-specific peptide	Insect	*Gastrophysa atrocyanea*	3	41	GCGGTCCGTATTGGTCCGTGCGACCAGGTTTGTC	58.9	0.90	GCCGTCCGTATTGGCCCGTGCGACCAGGTGTGCCC	59.7	0.92	GCCGTTCGTATTGGTCCGTGCGATCAAG	53.5	0.92	123	112.74	82.54	114.00

	AYCSGGG MYCN						CGCGCATC GTGCCGGA GCGCCACG AATGCTGTC GTGCACAT GGCCGTAG CGGCTACG CCTATTGTA GCGGTGGT GGTATGTAC TGCAAC			GCGTATTGT CCCGGAAC GTCACGAAT GCTGCCGT GCCCATGG CCGCTCAG GTTATGCGT ACTGCAGCG GCGGTGGC ATGTATTGT AAC			TTTGCCC GCGTATT GTGCCAG AACGCCA TGAATGT TGCCGCG CCCATGG CCGTAGC GGTTACG CATATTG CAGCGGC GGTGGTA TGTATTG CAAT						
S6	INKDCLLP MDVGRCR ASHPRYY YNSSSKR CEKFIYGG CRGNANN FHTLEECE KVCGVR	Kunitz-type proteinase inhibitor kalicludin-1	Sea anemone	*Anemonia sulcata*	3	58	ATTAACAAA GACTGTTTG CTGCCGAT GGATGTGG GTCGTTGTC GTGCGAGC CACCCGCG TTACTATTA CAATAGCA GCTCCAAA CGTTGCGA AAAGTTCAT CTATGGTG GCTGCCGC GGTAACGC AAATAACTT TCATACCCT GGAAGAGT GCGAGAAG GTTTGTGG CGTCCGC	48.9	0.86	ATCAATAAG GACTGTCTG CTGCCGATG GATGTGGG CCGTTGCCG TGCGTCTCA CCCGCGTTA CTACTACAA TTCGTCGTC AAAGCGTTG CGAAAAATT TATTTATGG CGGTTGTCG TGGCAACGC GAACAATTT CCATACCCT GGAAGAATG CGAAAAAGT GTGTGGTGT TCGC	47.8	0.88	ATCAACA AAGACTG CCTGCTT CCGATGG ATGTCGG CCGTTGC CGTGCCA GTCATCC ACGTTAC TACTATAA TTCCAGT AGCAAAC GTTGCGA GAAATTC ATCTATG GCGGTTG CCGTGGT AACGCGA ACAATTTC CATACCT TAGAAGA ATGTGAA AAAGTCT GCGGTGT GCGC	46.7	0.85	174	34.01	37.78	38.64
S7	TEEMPAL CHLQPDV PKCRGYF PRYYYNP EVGKCEQ FIYGGCG GNKNNFV SFEACRA TCIIPL	Kunitz-type proteinase inhibitor SHTX-3	Sea anemone	*Stichodactyla haddoni*	3	62	ACGGAAGA GATGCCGG CACTGTGTC ACTTGCAG CCGGACGT GCCAAAAT GTCGTGGC TACTTTCCG CGCTATTAC TATAACCCG GAAGTTGG TAAATGCGA GCAATTCAT CTACGGCG GTTGTGGT GGTAATAAG AATAACTTT GTCAGCTTC GAGGCCTG CCGTGCGA CCTGCATTA TCCCGCTG	50.5	0.83	ACCGAAGAA ATGCCGGCT CTGTGTCAC CTGCAACCG GATGTCCCG AAATGTCGT GGCTACTTC CCGCGTTAC TACTACAAC CCGGAAGT GGGCAAATG CGAACAGTT TATTTATGG CGGTTGTGG CGGTAACAA GAACAATTT TGTTAGCTT CGAAGCGT GCCGTGCC ACCTGTATT ATCCCGCTG	50.0	0.90	ACTGAAG AAATGCC AGCATTA TGCCATC TGCAGCC GGACGTG CCGAAGT GCCGTGG TTACTTTC CGCGTTA TTACTATA ATCCAGA AGTCGGC AAATGCG AACAATTT ATTTACG GCGGTTG CGGTGGC AACAAAA ATAATTTT GTTAGCT TTGAGGC CTGCCGT GCGACCT GCATTATT CCGCTG	46.4	0.86	186	22.55	22.43	27.97

S8	EDNCIAEDYGKCTWGGTKCCRGRPCRCSMIGTNCECTPRLIMEGLSFA	omega-agatoxin-Aa4b	Spider	*Agelenopsis aperta*	4	48	GAAGATAATTGTATCGCAGAGGACTACGGTAAGTGCACCTGGGGTGGTACGAAATGTTGCCGTGGCCGTCCGTGCCGTTGTAGCATGATCGGTACTAACTGCGAGTGCACCCCACGCCTGATTATGGAAGGCCTGAGCTTCGCG	53.3	0.86	GAAGATAACTGTATTGCTGAAGACTATGGTAAATGCACCTGGGGCGGCACCAAATGTTGTCGCGGTCGTCCGTGTCGTTGCTCTATGATTGGCACCAACTGCGAATGTACGCCGCGTCTGATCATGGAAGGTCTGAGCTTTGCG	50.0	0.94	GAGGACAATTGCATTGCCGAAGACTATGGTAAATGTACCTGGGGTGGCACTAAATGCTGTCGCGGCCGTCCATGTCGTTGTAGTATGATTGGCACCAATTGCGAATGCACCCCGCGCTTAATTATGGAGGGTCTGTCGTTTGCG	49.3	0.86	144	63.58	103.84	87.38
S9	GSCIESGKSCTHSRSMKNGLCCPKSRCNCRQIQHRHDYLGKRKYSCRCS	Omega-segestritoxin-Sf1a	Spider	*Segestria florentina*	4	49	GGCAGCTGCATCGAGAGCGGTAAAAGCTGCACCCACTCTCGTAGCATGAAGAATGGCCTGTGTTGTCCGAAATCCCGCTGCAACTGCCGCCAAATTCAGCACCGTCATGACTACCTGGGTAAACGTAAGTATAGCTGCCGTTGTAGC	51.6	0.88	GGCTCGTGTATTGAAAGCGGTAAATCCTGCACCCACTCCCGTAGCATGAAAAATGGTCTGTGTTGCCCGAAATCCCGTTGTAACTGCCGTCAGATTCAACATCGCCACGATTATCTGGGCAAGCGTAAGTACAGCTGCCGCTGTTCT	49.0	0.87	GGCTCGTGCATTGAAAGTGGTAAAAGCTGTACTCATAGTCGCAGCATGAAGAATGGCCTTTGCTGCCCGAAATCGCGTTGTAACTGCCGCCAGATTCAACACCGTCACGATTATTTAGGCAAACGCAAATACTCCTGTCGTTGTAGC	46.4	0.83	147	105.97	84.72	76.83
S10	LECHNQQSSQPPTTKSCPGDTNCYNKRWRDHRGTIIERGCGCPTVKPGINLKCCTTDRCNN	Short neurotoxin 1	Snake	*Hemachatus haemachatus*	4	61	CTGGAGTGCCATAATCAGCAGTCCAGCCAACCACCGACCACTAAGAGCTGTCCGGGTGATACCAACTGCTACAACAAGCGTTGGCGCGACCACCGTGGTACGATCATTGAACGTGGCTGTGGTTGCCCGACCGTTAAACCGGGCATCAATCTGAAATGCTGTACGACCGACCGCTGCAATAAC	53.4	0.86	CTGGAATGTCATAACCAACAAAGCTCGCAACCGCCGACGACGAAGTCATGTCCGGGTGATACCAACTGCTACAACAAACGCTGGCGCGATCATCGTGGCACCATTATCGAACGCGGCTGCGGTTGTCCGACGGTGAAACCGGGTATTAACCTGAAGTGCTGTACCACGGACCGTTGCAACAAT	52.4	0.87	CTTGAATGCCATAATCAGCAAAGCTCGCAGCCGCCGACCACTAAATCCTGCCCGGGCGATACCAATTGTTATAATAAACGTTGGCGCGACCATCGTGGCACTATCATTGAACGTGGTTGTGGTTGCCCGACCGTCAAACCGGGTATCAATCTGAAATGTTGTACC	48.7	0.88	183	100.08	86.13	118.66

													ACGGATC GCTGTAA CAAC					
S11	MTCYNQQ SSEAKTTT TCSGGVS SCYKKTW SDGRGTII ERGCGCP SVKKGIER ICCRTDKC NN	Short neurotoxin 1	Snake	*Oxyuranus scutellatus*	4	62	ATGACCTG CTATAATCA ACAGAGCT CCGAGGCG AAAACCAC GACCACGT GCAGCGGT GGTGTTAG CTCTTGTTA CAAGAAAAC TTGGAGCG ATGGTCGC GGCACGAT TATTGAGCG TGGTTGTG GCTGCCCG AGCGTGAA GAAGGGCA TCGAACGTA TCTGCTGTC GTACCGAC AAATGCAAC AAC	51.0	0.85	ATGACCTGC TACAACCAA CAATCGTCG GAAGCCAAG ACCACCACC ACCTGCTCG GGCGGCGT GAGTTCGTG CTACAAGAA GACCTGGA GCGATGGC CGTGGTACC ATTATCGAA CGCGGCTG CGGTTGTCC GTCTGTGAA AAAGGGCAT TGAACGTAT CTGCTGTCG CACGGACAA ATGCAACAA T	53.1	0.83	ATGACGT GTTATAAT CAGCAGT CCAGCGA AGCGAAA ACCACGA CGACGTG TTCGGGC GGTGTGA GTAGTTG TTATAAAA AAACTTG GAGTGAC GGTCGTG GCACTAT CATTGAG CGCGGCT GTGGTTG CCCGTCC GTTAAAA AAGGTAT TGAGCGT ATTTGTTG TCGCACC GATAAAT GTAATAAT	44.8	0.85	186	3.73	2.66	3.44
S12	RICLNQQ QSTPEDQ PTNGQCYI KTDCQNK TWNTHRG SRTDRGC GCPKVKP GINLRCCK TDKCNE	Short neurotoxin 1	Snake	*Bungarus fasciatus*	4	64	CGTATCTGT CTGAATCAA CAGCAGAG CACCCCGG AAGATCAAC CGACGAAT GGCCAATG CTACATTAA GACGGACT GCCAGAAC AAAACCTG GAACACCC ACCGCGGC AGCCGCAC TGACCGTG GTTGCGGT TGCCCAAA GGTTAAGC CGGGTATC AACCTGCG TTGCTGTAA AACCGATAA ATGTAATGA G	51.0	0.84	CGTATCTGC CTGAACCAA CAACAATCA ACCCCGGAA GACCAACCG ACCAACGGC CAATGCTAC ATCAAAACG GACTGTCAA AATAAGACC TGGAACACG CATCGTGGT AGCCGTACC GATCGTGGT TGCGGTTGT CCGAAAGTG AAGCCGGG TATTAACCT GCGCTGCT GTAAAACGG ACAAGTGCA ATGAA	49.5	0.83	CGTATTT GCCTTAA CCAGCAG CAGAGTA CTCCGGA AGATCAG CCAACTA ACGGCCA ATGCTAT ATTAAAAC CGATTGC CAGAACA AAACTTG GAACACC CATCGTG GTTCCCG TACGGAT CGCGGTT GCGGTTG CCCGAAA GTTAAGC CGGGTAT TAATCTG CGTTGCT GCAAAAC GGATAAG TGCAATG AA	47.0	0.86	192	94.88	86.95	84.26
S13	LTCVTSKS IFGITTENC PDGQNLC FKKWYYIV PRYSDITW GCAATCP KPTNVRE	Muscarinic toxin 1	Snake	*Dendroaspis angusticeps*	4	66	CTGACGTG TGTTACTAG CAAGAGCA TCTTTGGCA TCACGACC GAGAATTG CCCGGACG GTCAGAAC	50.0	0.84	CTGACCTGT GTGACCTCA AAATCTATC TTCGGTATT ACGACGGAA AACTGCCCG GACGGCCA GAACCTGTG	49.1	0.89	CTGACCT GTGTTAC TTCCAAG TCGATTTT TGGCATT ACTACGG AAAACTG TCCGGAT	46.6	0.86	198	17.31	24.28	21.71

O

	TIRCCETD KCNE						CTGTGCTTC AAGAAATG GTACTATAT TGTCCCGC GTTACTCCG ACATTACCT GGGGTTGT GCGGCAAC GTGTCCGA AACCAACCA ATGTGCGT GAAACCATC CGCTGCTG CGAAACCG ATAAATGCA ACGAG			CTTCAAAAA ATGGTATTA TATTGTGCC GCGTTACAG CGATATCAC GTGGGGTT GCGCAGCA ACCTGTCCG AAACCGACG AACGTTCGT GAAACCATT CGCTGCTGT GAAACCGAC AAGTGCAAT GAA			GGTCAGA ATCTTTG CTTTAAAA AATGGTA CTATATC GTGCCGC GTTATTC CGATATC ACCTGGG GCTGCGC GGCCACC TGCCCGA AACCGAC CAACGTG CGTGAAA CGATCCG TTGTTGT GAAACGG ATAAATGT AACGAG						
S14	MECYRCG VSGCHLKI TCSAEETF CYKWLNKI SNERWLG CAKTCTEI DTWNVYN KCCTTNL CNT	Bucandin	Snake	*Bungarus candidus*	5	63	ATGGAATGT TATCGCTGC GGCGTGTC CGGTTGTC ACCTGAAG ATTACGTGT AGCGCAGA AGAGACTTT CTGTTACAA ATGGCTGA ATAAGATCA GCAACGAG CGTTGGCT GGGTTGCG CGAAAACG TGCACCGA GATCGACA CCTGGAAT GTTTACAAC AAATGCTGC ACGACCAA CTTGTGCAA TACC	47.7	0.83	ATGGAATGT TATCGTTGC GGCGTCTC GGGCTGTCA CCTGAAGAT TACGTGCTC GGCAGAAG AAACCTTTT GCTACAAGT GGCTGAATA AAATTAGCA ACGAACGTT GGCTGGGC TGCGCGAA GACCTGTAC GGAAATCGA TACCTGGAA CGTGTATAA TAAATGCTG TACCACGAA CCTGTGCAA TACG	46.7	0.87	ATGGAAT GCTATCG CTGTGGT GTCTCCG GTTGCCA TCTTAAAA TCACCTG TAGTGCG GAAGAAA CTTTTTGC TATAAGT GGCTGAA TAAAATTT CGAATGA ACGTTGG CTTGGTT GTGCGAA AACGTGC ACGGAGA TCGATAC CTGGAAT GTGTATA ATAAGTG CTGTACC ACCAACC TGTGTAA CACG	42.6	0.84	189	0.72	0.09	0.97
S15	LTCKTCPF NTCANSE TCPAGKNI CYQKKWE EHRGERIE RRCVANC PKLGSND KSLLCCR RDDCN	Long neurotoxin MS4	Snake	*Micrurus surinamensis surinamensis*	5	64	CTGACCTG CAAGACCT GCCCGTTC AACACTTGT GCCAATAG CGAAACGT GTCCAGCG GGTAAGAAT ATCTGCTAC CAGAAGAA ATGGGAAG AGCACCGT GGTGAGCG CATTGAGC GTCGCTGC GTTGCAAAT TGTCCGAAA CTGGGCTC	51.5	0.82	CTGACCTGT AAGACCTGT CCGTTCAAT ACCTGTGCG AATAGCGAA ACCTGTCCG GCTGGCAAA AACATCTGC TACCAAAAG AAATGGGAA GAACATCGT GGCGAACG CATTGAACG TCGCTGCGT GGCGAACT GTCCGAAAC TGGGTAGCA ATGATAAGT	50.0	0.89	TTAACTTG TAAGACC TGTCCGT TTAACAC CTGTGCG AACAGTG AGACTTG TCCGGCA GGCAAAA ATATTTGC TATCAAAA AAAATGG GAAGAAC ACCGTGG TGAACGC ATTGAGC GTCGCTG TGTGGCA	45.5	0.85	192	49.76	54.71	42.72

							CAACGACA AAAGCTTGC TGTGTTGCC GTCGTGAC GATTGCAAC			CTCTGCTGT GCTGTCGTC GCGATGACT GCAAC			AATTGCC CAAAACT GGGCAGT AACGACA AAAGTCT GCTGTGC TGCCGCC GTGATGA TTGCAAT						
S16	RTCLISPS STPQTCP NGQDICFL KAQCDKF CSIRGPVI EQGCVAT CPQFRSN YRSLLCCT TDNCNH	Kappa-1-bungarotoxin	Snake	*Bungarus multicinctus*	5	66	CGCACCTG TCTGATTTC CCCGTCTA GCACTCCG CAGACGTG CCCGAACG GTCAAGAC ATCTGCTTT CTGAAAGC GCAATGCG ACAAGTTCT GTAGCATC CGTGGCCC AGTTATTGA GCAGGGTT GCGTGGCA ACCTGTCC GCAGTTCC GTAGCAATT ACCGTAGC TTGCTGTGT TGCACGAC CGATAACTG CAATCAC	52.9	0.84	CGCACCTGC CTGATTAGC CCGTCGTCA ACCCCGCAA ACCTGCCCG AATGGTCAA GACATCTGC TTCCTGAAA GCCCAGTGT GATAAATTT TGCAGTATT CGTGGCCC GGTGATCGA ACAGGGTTG CGTTGCGAC CTGTCCGCA ATTCCGTAG CAACTATCG CTCTCTGCT GTGCTGTAC CACGGATAA CTGTAATCA T	51.5	0.90	CGTACGT GTCTGAT CAGCCCA TCCAGCA CTCCGCA AACCTGT CCGAACG GCCAGGA TATTTGCT TTCTTAAA GCGCAAT GCGATAA GTTTTGC AGCATTC GCGGCCC GGTTATC GAACAGG GTTGCGT TGCAACT TGTCCAC AATTTCG CTCCAAT TATCGCT CGTTGCT GTGTTGT ACTACGG ACAACTG TAATCAT	48.0	0.80	198	82.09	73.13	87.03
S17	RECYLNP HDTQTCP SGQEICYV KSWCNA WCSSRGK VLEFGCA ATCPSVN TGTEIKCC SADKCNT YP	Long neurotoxin Ls3	Snake	*Laticauda semifasciata*	5	66	CGTGAATGT TACCTGAAT CCACACGA CACCCAGA CGTGTCCG AGCGGCCA AGAAATCTG CTACGTGAA GAGCTGGT GCAACGCG TGGTGCAG CTCTCGCG GCAAAGTC CTGGAGTT CGGTTGCG CCGCAACC TGCCCGAG CGTTAACAC TGGTACGG AGATTAAGT GTTGCTCC GCGGATAA ATGTAATAC CTATCCG	53.9	0.84	CGTGAATGT TACCTGAAT CCGCATGAT ACGCAAACC TGTCCGTCG GGTCAAGAA ATCTGCTAT GTGAAATCG TGGTGCAAC GCCTGGTG CAGCTCTCG TGGCAAAGT GCTGGAATT TGGTTGCGC GGCCACCT GTCCGAGTG TTAACACCG GCACGGAAA TTAAATGCT GTAGCGCG GATAAGTGT AATACGTAT CCG	52.0	0.89	CGCGAAT GCTACTT GAACCCA CATGATA CGCAGAC CTGCCCG AGCGGCC AAGAAAT CTGTTAT GTTAAAA GCTGGTG TAATGCG TGGTGTA GTAGTCG TGGCAAA GTGTTGG AGTTTGG TTGTGCC GCGACTT GCCCAAG CGTTAAC ACGGGTA CTGAGAT TAAGTGC TGTAGCG CAGATAA GTGTAAC	49.0	0.80	198	6.85	9.61	7.51

													ACGTACCCG					
S18	RTCLISPSSTSQTCPKGQDICFTKAFCDRWCSSRGPVIEQGCAATCPEFTSRYKSLLCCTTDNCNH	Kappa-flavitoxin	Snake	*Bungarus flaviceps*	5	66	CGTACCTGCTTGATCAGCCCGAGCAGCACCTCTCAGACCTGCCCAAAAGGTCAAGACATTTGCTTCACTAAGGCCTTCTGTGACCGCTGGTGTAGCTCCCGTGGCCCGGTTATCGAGCAGGGTTGTGCAGCGACGTGCCCGGAATTTACGAGCCGTTACAAAAGCCTGCTGTGTTGCACCACGGATAATTGCAACCAC	54.4	0.84	CGCACCTGCCTGATTTCCCCGTCCAGCACCTCCCAGACCTGTCCGAAGGGCCAAGATATTTGTTTACCAAGGCGTTTTGTGACCGCTGGTGCAGCTCTCGTGGCCCGGTGATTGAACAGGGTTGCGCGGCCACCTGTCCGGAATTTACGAGTCGCTATAAAGCCTGCTGTGCTGTACCACGGATAACTGTAATCAT	53.4	0.89	CGCACCTGCCTGATCTCCGCCGTCCTCGAGAGTCAGACCTGCCCAAAAGGTCAGGACATCTGCTTTACTAAAGCATTTTGTGACCGTTGGTGTAGCAGTCGTGGCCCAGTCATTGAACAGGGCTGTGCAGCGACGTGTCCAGAATTCACGAGTCGTTATAAAAGTCTGCTGTGCTGTACCACTGATAACTGCAACCAC	51.0	0.82	198	66.90	48.11	111.89
S19	KTCLKTPSSTPQTCPQGQDICFLKVSCEQFCPIRGPVIEQGCAATCPEFRSNDRSLLCCTTDNCNH	Kappa-2-bungarotoxin	Snake	*Bungarus multicinctus*	5	66	AAAACGTGTCTGAAAACTCCGAGCTCTACCCCGCAGACCTGCCCGCAAGGTCAAGACATCTGCTTCCTGAAGGTTTCCTGTGAGCAGTTCTGCCCGATTCGTGGTCCAGTGATCGAACAGGGCTGTGCAGCGACCTGTCCGGAGTTTCGTAGCAATGATCGCAGCTTGCTGTGCTGCACCACGGACAACTGCAACCAC	54.4	0.83	AAAACGTGTCTGAAAACCCCGTCATCAACGCCGCAAACCTGTCCGCAGGGCCAAGATATTTGCTTTCTGAAGGTGTCGTGTGAACAATTTTGCCCGATTCGTGGTCCGGTGATCGAACAGGGTTGCGCAGCAACCTGTCCGGAATTCCGTAGCAACGATCGCTCTCTGCTGTGCTGTACCACGGACAACTGTAATCAT	50.5	0.87	AAGACGTGTCTGAAGACCCCGTCGAGCACGCCACAGACCTGTCCGCAGGGCCAGGATATTTGTTTCCTGAAAGTTAGCTGCGAACAATTTTGTCCAATCCGTGGTCCGGTTATCGAACAGGGTTGTGCGGCGACGTGTCCGGAATTTCGTTCCAACGACCGTAGTCTGTTATGCTGCACCACCGATAACTGCAACCAC	52.0	0.84	198	46.19	62.58	54.05
S20	IVCHTTATSPISAVTCPPGENLCYRKMCDAICSSRGKVVELGCAATCPSKKP	Alpha-bungarotoxin N3	Snake	*Bungarus candidus*	5	73	ATTGTGTGTCACACGACCGCCACCAGCCCGATCTCCGCAGTTACCTGTCCGCCAGGTG	55.6	0.83	ATTGTCTGCCATACCACCGCAACCTCGCCGATCTCCGCCGTTACCTGTCCGCCGGTGAAAAC	54.2	0.90	ATTGTTTGTCACACGACGGCAACCAGCCCGATCAGTGCAGTGACTTGCCC	52.0	0.83	219	59.11	49.85	58.14

	YEEVTCC SNDKCNP HPKQRPG						AAAATCTGT GTTACCGC AAGATGTG CGATGCGA TCTGCAGC AGCCGTGG CAAAGTTGT CGAGCTGG GTTGCGCT GCGACGTG CCCGTCTAA AAAGCCGT ATGAAGAG GTGACTTG CTGTAGCAA CGACAAAT GCAACCCG CATCCGAA GCAGCGTC CTGGC			CTGTGCTAC CGCAAAATG TGTGATGCG ATTTGCAGC TCTCGTGGC AAAGTGGTT GAACTGGGT TGCGCGGC CACCTGTCC GAGTAAAAA GCCGTATGA AGAAGTCAC GTGCTGTAG CAACGATAA ATGTAATCC GCATCCGAA GCAGCGCC CGGGC			GCCAGGT GAAAACC TGTGTTA CCGTAAG ATGTGCG ACGCAAT TTGTAGT AGCCGCG GTAAAGT GGTCGAA TTGGGCT GTGCCGC GACCTGC CCGAGCA AAAAGCC ATACGAA GAAGTTA CCTGCTG CTCGAAT GATAAAT GTAACCC GCATCCG AAACAGC GTCCAGG C						
S21	AVITGACE RDLQCGK GTCCAVS LWIKSVRV CTPVGTS GEDCHPA SHKIPFSG QRMHHTC PCAPNLA CVQTSPK KFKCLSKS	MIT1	Snake	*Dendroaspis polylepis polylepis*	5	81	GCGGTGAT TACGGGTG CGTGTGAA CGTGACTT GCAATGTG GTAAAGGC ACCTGCTGT GCGGTCAG CCTGTGGA TCAAGAGC GTTCGCGTT TGCACCCC GGTGGGTA CCTCTGGT GAGGATTG CCATCCGG CTAGCCAC AAAATCCCG TTTAGCGG CCAGCGTA TGCATCACA CTTGCCCG TGTGCCCC AAACCTGG CATGCGTC CAGACGAG CCCGAAAA AGTTCAAGT GCCTGTCC AAAAGC	56.2	0.84	GCGGTGATT ACGGGTGC GTGTGAACG TGACCTGCA ATGTGGCAA GGGTACCTG CTGTGCGGT GTCTCTGTG GATTAAGTC GGTGCGTGT GTGCACCCC GGTTGGCAC GAGCGGTG AAGATTGTC ATCCGGCGA GTCACAAAA TTCCGTTTT CCGGCCAG CGTATGCAT CACACCTGC CCGTGTGCA CCGAACCTG GCATGCGTC CAAACGAGC CCGAAAAAG TTCAAATGT CTGAGCAAG TCT	55.0	0.85	GCGGTTA TCACGGG CGCCTGC GAACGCG ACCTGCA GTGCGGT AAAGGTA CGTGTTG CGCAGTC AGTCTGT GGATCAA AAGCGTG CGTGTTT GTACCCC AGTTGGT ACCTCCG GCGAAGA TTGCCAT CCAGCCA GTCACAA AATTCCG TTTAGCG GTCAACG TATGCAT CATACCT GTCCATG CGCACCG AACCTGG CGTGCGT GCAGACC TCCCCGA AAAAATTT AAGTGCC TTAGCAA AAGC	53.4	0.86	243	2.71	1.95	3.78
S22	ACIPRGEI CTDDCEC CGCDNQC YCPPGSS LGIFKCSC	Omega-ctenitoxin-Pn4a	Spider	*Phoneutria nigriventer*	6	55	GCCTGCATT CCGCGTGG TGAGATCTG CACCGACG ATTGCGAGT	50.3	0.84	GCGTGTATT CCGCGTGG TGAAATCTG TACCGATGA CTGTGAATG	46.2	0.84	GCCTGCA TTCCACG TGGTGAA ATTTGTAC GGATGAT	44.4	0.90	165	67.46	80.19	72.03

	AHANKYF CNRKKEK CKKA						GTTGTGGC TGTGACAAC CAGTGTTAT TGCCCACC GGGTAGCT CCCTGGGC ATCTTCAAA TGCAGCTG TGCGCACG CAAACAAAT ACTTTTGCA ATCGCAAG AAAGAAAAG TGCAAGAAA GCG			TTGTGGCTG CGATAATCA ATGTTACTG TCCGCCGG GTTCGTCGC TGGGCATTT TTAAATGCA GCTGTGCG CATGCCAAC AAGTACTTC TGCAACCGT AAAAAGGAA AAGTGTAAA AAGGCG			TGCGAAT GTTGTGG CTGTGAT AATCAGT GCTACTG TCCGCCG GGCAGTA GTCTGGG CATTTTTA AATGTTC CTGCGCG CATGCGA ACAAATA CTTCTGT AACCGCA AAAAAGA GAAATGC AAAAAAG CC						
S23	SCIDIGGD CDGEKDD CQCCRRN GYCSCYS LFGYLKSG CKCVVGT SAEFQGIC RRKARQC YNSDPDK CESHNKP KRR	omega-agatoxin-Aa3a	Spider	*Agelenopsis aperta*	6	76	AGCTGCAT CGATATTGG TGGTGACT GTGACGGT GAGAAAGA CGACTGTC AGTGCTGC CGTCGTAA CGGTTATTG TTCCTGCTA CAGCCTGTT CGGCTATCT GAAGAGCG GTTGTAAAT GCGTTGTG GGCACCTC TGCAGAATT TCAAGGCAT CTGCCGTC GCAAAGCG CGCCAGTG TTACAATAG CGATCCGG ATAAGTGC GAGAGCCA CAACAAGC CGAAACGT CGC	51.7	0.86	TCCTGTATT GATATTGGC GGTGATTGT GATGGCGAA AAAGATGAT TGTCAGTGC TGTCGTCGT AACGGCTAT TGCTCGTGC TACTCCCTG TTTGGCTAT CTGAAAAGT GGTTGCAAG TGTGTGGTT GGCACCAG CGCGGAATT CCAGGGTAT TTGCCGTCG CAAAGCCCG TCAATGTTA CAACAGCGA TCCGGACAA GTGCGAATC TCATAATAA ACCGAAGC GTCGC	48.3	0.89	TCGTGCA TTGACATT GGTGGCG ATTGTGA TGGCGAG AAGGATG ATTGTCA GTGTTGC CGTCGTA ACGGCTA TTGTAGC TGCTATA GTCTTTTT GGCTATC TTAAAAG CGGTTGC AAATGCG TTGTGGG TACGAGC GCAGAGT TCCAAGG CATTTGT CGTCGTA AAGCACG TCAATGC TATAATTC CGATCCG GATAAAT GCGAGTC GCACAAT AAACCGA AGCGTCG T	46.2	0.83	228	8.76	8.93	7.45
S24	HPTKPCM YCSFGQC VGPHICC GPTGCEM GTAEANM CSEEDED PIPCQVFG SDCALNN PDNIHGH CVADGICC VDDTCTT HLGCL	Conophysin-R	Cone snail	*Conus radiatus*	7	84	CATCCGAC CAAACCGT GCATGTACT GCAGCTTC GGTCAATGT GTTGGTCC GCACATTTG TTGTGGTCC GACGGGTT GCGAGATG GGCACCGC CGAAGCGA ATATGTGTA	55.0	0.85	CATCCGACG AAACCGTGT ATGTACTGT AGCTTTGGT CAATGTGTT GGTCCGCAT ATCTGCTGT GGTCCGAC GGGCTGTG AAATGGGCA CCGCGGAA GCCAACATG TGCAGCGAA	51.9	0.90	CACCCAA CTAAACC GTGTATG TATTGCA GCTTTGG TCAGTGT GTCGGTC CGCATAT TTGCTGC GGTCCGA CCGGCTG CGAAATG GGCACGG	49.2	0.83	252	33.26	42.05	27.56

							GCGAAGAG GATGAGGA TCCGATCC CGTGCCAG GTGTTTGG CTCCGACT GCGCGCTG AATAACCCA GACAACATC CACGGCCA TTGTGTCGC AGACGGCA TTTGCTGCG TTGATGACA CCTGTACG ACTCACTTG GGTTGCCT G			GAAGATGAA GACCCGATT CCGTGCCA GGTCTTTGG CTCTGATTG TGCACTGAA CAATCCGGA CAATATTCA TGGCCACTG TGTGGCGG ATGGTATCT GCTGTGTTG ATGACACGT GCACCACG CATCTGGGT TGTCTG			CCGAAGC CAATATG TGCTCGG AAGAAGA TGAAGAT CCAATCC CATGCCA GGTTTTT GGTTCGG ATTGTGC GTTAAAC AACCCAG ATAACATT CATGGCC ACTGTGT GGCGGAT GGCATTT GCTGTGT TGATGAT ACTTGTA CTACTCA CCTTGGC TGTCTC						

Table S5. 2| Properties of the novel prokaryotic expression vectors.

Vector short name	Expression vector	Fusion Protein	Fusion protein sequence	Tag size (nt)	MW tag (kDa)	Tag position	Promoter	Resistance	Notes
pHTP1 (His)	pHTP-His	MGSS-His6-SSGPQQGLR	MGSSHHHHHSSGPQQGLR	57	1.9	N-terminal	T7/lac	Kan	-
pHTP2 (LLDsbC)	pHTP-LLDsbC	Leader less Disulfide-bond isomerase DsbC-MGSS-His6-SSGPQQGLR	MGDDAAIQQTLAKMGIKSSDIQPAPVAGMKT VLTNSGVLYITDDGKHIIQGPMYDVSGTAPV NVTNKMLLKQLNALEKEMIVYKAPQEKHVIT VFTDITCGYCHKLHEQMADYNALGITVRYLA FPRQGLDSDAEKEMKAIWCAKDKNKAFDDV MAGKSVAPASCDVDIADHYALGVQLGVSGT PAVVLSNGTLVPGYQPPKEMKEFLDEHQKM TSGKGSSMGSSHHHHHHSSGPQQGLR	720	26.0	N-terminal	T7/lac	Kan	-
pHTP3 (mutDsbC)	pHTP-mutDsbC	Mutant of leader less Disulfide-bond isomerase DsbC-MGSS-His6-SSGPQQGLR	MGDDAAIQQTLAKMGIKSSDIQPAPVAGMKT VLTNSGVLYITDDGKHIIQGPMYDVSGTAPV NVTNKMLLKQLNALEKEMIVYKAPQEKHVIT VFTDITAGYAHKLHEQMADYNALGITVRYLA FPRQGLDSDAEKEMKAIWCAKDKNKAFDDV MAGKSVAPASCDVDIADHYALGVQLGVSGT PAVVLSNGTLVPGYQPPKEMKEFLDEHQKM TSGKGSSMGSSHHHHHHSSGPQQGLR	720	25.9	N-terminal	T7/lac	Kan	tag contains two mutations: Cys100Ala and Cys103Ala
pHTP4 (DsbC)	pHTP-DsbC	Disulfide-bond isomerase DsbC-MGSS-His6-SSGPQQGLR	MGKKGFMLFTLLAAFSGFAQADDAAIQQTLA KMGIKSSDIQPAPVAGMKTVLTNSGVLYITD DGKHIIQGPMYDVSGTAPVNVTNKMLLKQL NALEKEMIVYKAPQEKHVITVFTDITCGYCHK LHEQMADYNALGITVRYLAFPRQGLDSDAE KEMKAIWCAKDKNKAFDDVMAGKSVAPASC DVDIADHYALGVQLGVSGTPAVVLSNGTLVP GYQPPKEMKEFLDEHQKMTSGKGSSMGSS HHHHHHSSGPQQGLR	777	28.0	N-terminal	T7/lac	Kan	includes a signal peptide (KKGFMLFTLLAAFSGFAQA)
pHTP5 (LLMBP)	pHTP-LLMBP	Leader less Maltose binding protein-MGSS-His6-SSGPQQGLR	MGKIEEGKLVIWINGDKGYNGLAEVGKKFEK DTGIKVTVEHPDKLEEKFPQVAATGDGPDIIF WAHDRFGGYAQSGLLAEITPDKAFQDKLYP FTWDAVRYNGKLIAYPIAVEALSLIYNKDLLP NPPKTWEEIPALDKELKAKGKSALMFNLQEP YFTWPLIAADGGYAFKYENGKYDIKDVGVD NAGAKAGLTFLVDLIKNKHMNADTDYSIAEA AFNKGETAMTINGPWAWSNIDTSKVNYGVT VLPTFKGQPSKPFVGVLSAGINAASPNKELA KEFLENYLLTDEGLEAVNKDKPLGAVALKSY EEELAKDPRIAATMENAQKGEIMPNIPQMSA FWYAVRTAVINAASGRQTVDEALKDAQTSM GSSHHHHHHSSGPQQGLR	1164	42.6	N-terminal	T7/lac	Kan	-
pHTP6 (MBP)	pHTP-MBP	Maltose binding protein-MGSS-His6-SSGPQQGLR	MGKKGFMLFTLLAAFSGFAQAKIEEGKLVIWI NGDKGYNGLAEVGKKFEKDTGIKVTVEHPD KLEEKFPQVAATGDGPDIIFWAHDRFGGYA QSGLLAEITPDKAFQDKLYPFTWDAVRYNG KLIAYPIAVEALSLIYNKDLLPNPPKTWEEIPA LDKELKAKGKSALMFNLQEPYFTWPLIAADG GYAFKYENGKYDIKDVGVDNAGAKAGLTFL VDLIKNKHMNADTDYSIAEAAFNKGETAMTI NGPWAWSNIDTSKVNYGVTVLPTFKGQPSK PFVGVLSAGINAASPNKELAKEFLENYLLTD EGLEAVNKDKPLGAVALKSYEEELAKDPRIA ATMENAQKGEIMPNIPQMSAFWYAVRTAVI NAASGRQTVDEALKDAQTSMGSSHHHHHH SSGPQQGLR	1221	44.6	N-terminal	T7/lac	Kan	includes a signal peptide (KKGFMLFTLLAAFSGFAQA)

Table S5. 3| Properties of peptides selected for identification of the best expression vector to produce recombinant venom peptides in *E. coli*.

Peptide	Primary sequence	Peptide name	Organism	Disulfide-bridges	Peptide size (aa)	Gene sequence	GC content (%)	CAI	Gene size (nt)
T1	AEKDCIAPGAPCFGTDKPCCN PRAWCSSYANKCL	Ptu1	*Peirates turpis*	3	34	GCCGAGAAAGACTGCATCGCTCCGGG TGCGCCGTGTTTCGGCACCGATAAGC CGTGCTGTAATCCACGTGCGTGGTGT AGCAGCTATGCGAATAAATGTCTG	58.8	0.86	102
T2	AVRIGPCDQVCPRIVPERHECC RAHGRSGYAYCSGGGMYCN	Diapause-specific peptide	*Gastrophysa atrocyanea*	3	41	GCCGTTCGTATTGGTCCGTGCGACCA GGTTTGTCCGCGCATCGTGCCGGAGC GCCACGAATGCTGTCGTGCACATGGC CGTAGCGGCTACGCCTATTGTAGCGG TGGTGGTATGTATTGCAAT	63.1	0.9	123
T3	IDTCRLPSDRGRCKASFERWY FNGRTCAKFIYGGCGGNGNKF PTQEACMKRCAKA	Kappa-theraphotoxin-Hh1a	*Haplopelma schmidti*	3	55	ATTGATACGTGTCGCCTGCCGAGCGA CCGTGGTCGTTGCAAGGCTAGCTTCG AGCGTTGGTACTTCAATGGCCGCACC TGTGCAAAGTTCATCTACGGTGGTTGC GGTGGCAACGGCAACAAATTTCCGAC CCAAGAAGCCTGCATGAAACGTTGTG CGAAAGCG	55.3	0.86	165
T4	CTTGPCCRQCKLKPAGTTCWK TSLTSHYCTGKSCDCPLYPG	Short disintegrin obtustatin	*Macrovipera lebetina obtusa*	4	41	TGTACCACCGGCCCGTGCTGCCGCCA ATGCAAACTGAAGCCAGCCGGCACTA CGTGTTGGAAAACTTCGTTAACCAGCC ACTATTGTACTGGTAAGTCCTGTGATT GCCCGCTGTACCCAGGT	51.0	0.76	123
T5	EDNCIAEDYGKCTWGGTKCCR GRPCRCSMIGTNCECTPRLIM EGLSFA	omega-agatoxin-Aa4b	*Agelenopsis aperta*	4	48	GAGGACAATTGCATTGCAGAGGACTA CGGTAAGTGCACCTGGGGTGGTACGA AATGTTGCCGTGGCCGTCCGTGCCGT TGTAGCATGATCGGTACTAACTGCGA GTGCACCCCACGCCTGATTATGGAGG GTCTGTCGTTTGCG	57.6	0.82	144
T6	LTCVKSNSIWFPTSEDCPDGQ NLCFKRWQYISPRMYDFTRGC AATCPKAEYRDVINCCGTDKC NK	MT7	*Dendroaspis angusticeps*	4	65	CTGACGTGTGTCAAGAGCAATAGCAT CTGGTTCCCGACGAGCGAAGACTGCC CGGACGGTCAAAATCTGTGTTTCAAG CGTTGGCAATATATTAGCCCGCGTATG TACGATTTTACCCGTGGTTGCGCAGC AACGTGTCCGAAAGCGGAATATCGTG ATGTGATCAACTGCTGTGGTACCGAC AAATGCAATAAG	50.5	0.84	195
T7	NQVMTIFLVLGVIVYSVESSSTP DGTWVKCRHDCFTKYKSCQM SDSCHDEQSCHQCHVKHTDC VNTGCP	Acrorhagin-1	*Actinia equina*	4	69	AATCAAGTTATGACGATCTTCCTGGTG CTGGGTGTTATTGTTTACTCGGTCGAA TCCTCCTCAACGCCGGACGGTACCTG GGTGAAATGCCGTCACGATTGTTTTAC CAAATATAAGAGTTGCCAGATGAGCG ATTCTTGTCATGACGAACAGAGCTGCC ATCAATGTCACGTGAAACATACCGACT GCGTTAACACGGGCTGTCCG	48.2	0.84	207
T8	AWKCLPKDSTCGDDCDCCEG LHCHCPLRNMLPAILRCSCQSK DDHINTCPKYKKS	Omega-oxotoxin-Ol1b	*Oxyopes lineatus*	5	55	GCGTGGAAATGCTTACCGAAAGATTC GACGTGCGGCGACGATTGTGACTGCT GTGAAGGTTTGCACTGCCATTGTCCG CTGCGTAACATGCTCCCGGCAATCCT GCGTTGCTCCTGTCAGAGTAAGGATG ATCACATTAATACGTGCCCGAAATATA AAAAATCG	53.1	0.82	165

T9	RECYLNPHDTQTCPSGQEICY VKSWCNAWCSSRGKVLEFGC AATCPSVNTGTEIKCCSADKCN TYP	Long neurotoxin Ls3	*Laticauda semifasciata*	5	66	CGCGAATGCTACCTGAATCCACACGA CACCCAGACGTGTCCGAGCGGCCAAG AAATCTGCTACGTGAAGAGCTGGTGC AACGCGTGGTGCAGCTCTCGCGGCAA AGTCCTGGAGTTCGGTTGCGCCGCAA CCTGCCCGAGCGTTAACACTGGTACG GAGATTAAGTGTTGCTCCGCGGATAA GTGTAACACGTACCCG	59.2	0.81	198
T10	MKCKICNFDTCRAGELKVCAS GEKYCFKESWREARGTRIERG CAATCPKGSVYGLYVLCCTTD DCN	Candoxin	*Bungarus candidus*	5	66	ATGAAATGCAAGATCTGCAACTTCGAC ACCTGTCGCGCTGGCGAATTGAAGGT TTGCGCGAGCGGCGAAAAGTATTGTT TCAAAGAGAGCTGGCGTGAAGCCCGT GGTACCCGTATTGAGCGCGGTTGCGC GGCAACGTGTCCGAAAGGTTCCGTGT ACGGTCTGTATGTCCTGTGTTGCACG ACCGACGATTGCAAC	56.7	0.83	198
T11	AVITGACERDLQCGKGTCCAV SLWIKSVRVCTPVGTSGEDCH PASHKIPFSGQRMHHTCPCAP NLACVQTSPKKFKCLSKS	MIT1	*Dendroaspis polylepis polylepis*	5	81	GCTGTTATTACGGGTGCGTGTGAACG TGATCTGCAATGTGGCAAGGGTACCT GTTGTGCTGTCTCTCTGTGGATTAAGT CAGTGCGTGTGTGCACCCCGGTTGGC ACGAGCGGTGAAGATTGTCATCCGGC GAGTCACAAAATTCCGTTTTCCGGCCA GCGTATGCATCACACCTGCCCGTGTG CACCGAACCTGGCATGCGTCCAAACG AGCCCGAAAAAGTTCAAATGTCTGAG CAAGTCT	55.7	0.83	243
T12	EMCNMCVRPYPFMSSCCPEG QDRCYKSYWVNENGKQKKYH GKYPVILERGCVTACTGPGSG SIYNLYTCCPTNRCGSSSTSG	Muscarinic toxin BM14	*Bungarus multicinctus*	5	82	GAGATGTGCAACATGTGCGTGCGTCC GTACCCGTTCATGAGCAGCTGCTGTC CGGAAGGTCAGGACCGCTGCTATAAA TCCTATTGGGTTAACGAGAACGGCAA ACAGAAGAAATACCACGGCAAGTACC CGGTTATCCTGGAACGTGGTTGTGTC ACGGCGTGCACGGGTCCAGGTAGCG GCAGCATTTACAATTTGTATACCTGTT GCCCGACCAATCGTTGTGGCAGCAGC TCTACCAGCGGT	53.1	0.86	246
T13	ECDCGSPANPCCDAATCKLRP GAQCADGLCCDQCRFIKKGTV CRPARGDWNDDTCTGQSADC PRNGLYG	Disintegrin crotatroxin-4	*Crotalus atrox*	6	69	GAATGTGATTGCGGTAGCCCAGCCAA TCCGTGCTGCGACGCTGCGACGTGCA AACTGCGTCCGGGTGCGCAATGTGCA GATGGCCTGTGCTGTGACCAGTGCCG CTTTATCAAGAAAGGTACCGTTTGCCG CCCGGCACGTGGCGACTGGAACGAC GATACCTGTACCGGTCAAAGCGCGGA TTGTCCGCGTAATGGTCTGTACGGC	60.7	0.87	207
T14	SCIDIGGDCDGEKDDCQCCRR NGYCSCYSLFGYLKSGCKCVV GTSAEFQGICRRKARQCYNSD PDKCESHNKPKRR	omega-agatoxin-Aa3a	*Agelenopsis aperta*	6	76	TCGTGCATTGACATTGGTGGTGACTGT GACGGTGAGAAAGACGACTGTCAGTG CTGCCGTCGTAACGGTTATTGTTCCTG CTACAGCCTGTTCGGCTATCTGAAGA GCGGTTGTAAATGCGTTGTGGGCACC TCTGCAGAATTTCAAGGCATCTGCCGT CGCAAAGCGCGCCAGTGTTACAATAG CGATCCGGATAAGTGCGAGAGCCACA ATAAACCGAAGCGTCGT	52.6	0.85	228
T15	SPPVCGNKILEQGEDCDCGSP ANCQDRCCNAATCKLTPGSQC	Disintegrin bitan	*Bitis arietans*	7	83	AGCCCGCCAGTGTGTGGCAACAAAAT CCTGGAACAGGGTGAGGACTGTGATT	54.2	0.86	249

	NYGECCDQCRFKKAGTVCRIA RGDWNDDYCTGKSSDCPWNH					GTGGTAGCCCGGCAAACTGTCAAGAT CGCTGCTGCAATGCGGCTACGTGCAA ACTGACCCCGGGTAGCCAGTGCAATT ATGGCGAATGTTGCGATCAATGCCGT TTTAAGAAGGCGGGCACCGTTTGTCG TATTGCCCGTGGTGACTGGAACGATG ACTACTGCACCGGTAAATCCTCTGACT GCCCGTGGAACCAC			
T16	HPTKPCMYCSFGQCVGPHICC GPTGCEMGTAEANMCSEEDE DPIPCQVFGSDCALNNPDNIHG HCVADGICCVDDTCTTHLGCL	Conophysin-R	*Conus radiatus*	7	84	CACCCGACCAAACCGTGTATGTACTG CAGCTTCGGCCAATGTGTTGGTCCGC ATATCTGCTGCGGCCCAACGGGCTGC GAGATGGGTACCGCGGAAGCCAACAT GTGTTCCGAAGAGGACGAAGATCCGA TTCCGTGCCAGGTTTTCGGTAGCGAC TGCGCACTGAACAATCCTGACAATATT CATGGTCACTGTGTCGCGGATGGCAT CTGCTGCGTGGACGATACCTGTACTA CGCACTTGGGTTGTCTG	56.1	0.84	252

In italic are represented the eight peptides used in the TEV optimization experiments.

Table S5. 4| Sequence characteristics of human Kin17 (KN17) protein and primer sequences used to produce 20 protein variants with different N-terminal amino acids.

Protein of interest	Amino acid	Codon	Primary protein sequence	Protein size (aa)	Gene sequence	Gene size (nt)	Forward primer (5´-3´)
KN17_A	Alanine (A)	GCG	AEEEKKRTARTDYWLQPEIIVKIITKKLGEK YHKKKAIVKEVIDKYTAVVKMIDSGDKLKL DQTHLETVIPAPGKRILVLNGGYRGNEGTL ESINEKTFSATIVIETGPLKGRRVEGIQYED ISKL	126	GCGGAAGAGGAGAAGAAAAGAACTGCCCGAACAGACT ACTGGCTACAGCCTGAAATTATTGTGAAAATTATAACCA AGAAACTGGGAGAGAAATATCATAAGAAAAAGGCTATT GTTAAGGAAGTAATTGACAAATATACAGCTGTTGTGAAG ATGATTGATTCTGGAGACAAGCTGAAACTTGACCAGAC TCATTTAGAGACAGTAATTCCAGCACCAGGAAAAAGAAT TCTAGTTTTAAATGGAGGCTACAGAGGAAATGAAGGTA CCCTAGAATCCATCAATGAGAAGACTTTTTCAGCTACTA TCGTCATTGAAACTGGCCCTTTAAAAGGACGCAGAGTT GAAGGAATTCAATATGAAGACATCTCTAAACTT	378	GGGGACAAGTTTGTACAAAAA AGCAGGCTTAGAAAACCTGTA CTTCCAGGCGGAAGAGGAGA AGAAAAGAAC
KN17_C	Cysteine (C)	TGC	CEEEKKRTARTDYWLQPEIIVKIITKKLGEK YHKKKAIVKEVIDKYTAVVKMIDSGDKLKL DQTHLETVIPAPGKRILVLNGGYRGNEGTL ESINEKTFSATIVIETGPLKGRRVEGIQYED ISKL	126	TGCGAAGAGGAGAAGAAAAGAACTGCCCGAACAGACTA CTGGCTACAGCCTGAAATTATTGTGAAAATTATAACCAA GAAACTGGGAGAGAAATATCATAAGAAAAAGGCTATTG TTAAGGAAGTAATTGACAAATATACAGCTGTTGTGAAGA TGATTGATTCTGGAGACAAGCTGAAACTTGACCAGACT CATTTAGAGACAGTAATTCCAGCACCAGGAAAAAGAATT CTAGTTTTAAATGGAGGCTACAGAGGAAATGAAGGTAC CCTAGAATCCATCAATGAGAAGACTTTTTCAGCTACTAT CGTCATTGAAACTGGCCCTTTAAAAGGACGCAGAGTTG AAGGAATTCAATATGAAGACATCTCTAAACTT	378	GGGGACAAGTTTGTACAAAAA AGCAGGCTTAGAAAACCTGTA CTTCCAGTGCGAAGAGGAGA AGAAAAGAAC
KN17_D	Aspartate (D)	GAT	DEEEKKRTARTDYWLQPEIIVKIITKKLGEK YHKKKAIVKEVIDKYTAVVKMIDSGDKLKL DQTHLETVIPAPGKRILVLNGGYRGNEGTL ESINEKTFSATIVIETGPLKGRRVEGIQYED ISKL	126	GATGAAGAGGAGAAGAAAAGAACTGCCCGAACAGACTA CTGGCTACAGCCTGAAATTATTGTGAAAATTATAACCAA GAAACTGGGAGAGAAATATCATAAGAAAAAGGCTATTG TTAAGGAAGTAATTGACAAATATACAGCTGTTGTGAAGA TGATTGATTCTGGAGACAAGCTGAAACTTGACCAGACT CATTTAGAGACAGTAATTCCAGCACCAGGAAAAAGAATT CTAGTTTTAAATGGAGGCTACAGAGGAAATGAAGGTAC CCTAGAATCCATCAATGAGAAGACTTTTTCAGCTACTAT CGTCATTGAAACTGGCCCTTTAAAAGGACGCAGAGTTG AAGGAATTCAATATGAAGACATCTCTAAACTT	378	GGGGACAAGTTTGTACAAAAA AGCAGGCTTAGAAAACCTGTA CTTCCAGGATGAAGAGGAGAA GAAAAGAAC
KN17_E	Glutamate (E)	GAA	EEEEKKRTARTDYWLQPEIIVKIITKKLGEK YHKKKAIVKEVIDKYTAVVKMIDSGDKLKL DQTHLETVIPAPGKRILVLNGGYRGNEGTL ESINEKTFSATIVIETGPLKGRRVEGIQYED ISKL	126	GAAGAAGAGGAGAAGAAAAGAACTGCCCGAACAGACTA CTGGCTACAGCCTGAAATTATTGTGAAAATTATAACCAA GAAACTGGGAGAGAAATATCATAAGAAAAAGGCTATTG TTAAGGAAGTAATTGACAAATATACAGCTGTTGTGAAGA TGATTGATTCTGGAGACAAGCTGAAACTTGACCAGACT CATTTAGAGACAGTAATTCCAGCACCAGGAAAAAGAATT CTAGTTTTAAATGGAGGCTACAGAGGAAATGAAGGTAC CCTAGAATCCATCAATGAGAAGACTTTTTCAGCTACTAT CGTCATTGAAACTGGCCCTTTAAAAGGACGCAGAGTTG AAGGAATTCAATATGAAGACATCTCTAAACTT	378	GGGGACAAGTTTGTACAAAAA AGCAGGCTTAGAAAACCTGTA CTTCCAGGAAGAAGAGGAGA AGAAAAGAAC
KN17_F	Phenilalanine (F)	TTT	FEEEKKRTARTDYWLQPEIIVKIITKKLGEK YHKKKAIVKEVIDKYTAVVKMIDSGDKLKL DQTHLETVIPAPGKRILVLNGGYRGNEGTL ESINEKTFSATIVIETGPLKGRRVEGIQYED ISKL	126	TTTGAAGAGGAGAAGAAAAGAACTGCCCGAACAGACTA CTGGCTACAGCCTGAAATTATTGTGAAAATTATAACCAA GAAACTGGGAGAGAAATATCATAAGAAAAAGGCTATTG TTAAGGAAGTAATTGACAAATATACAGCTGTTGTGAAGA TGATTGATTCTGGAGACAAGCTGAAACTTGACCAGACT CATTTAGAGACAGTAATTCCAGCACCAGGAAAAAGAATT CTAGTTTTAAATGGAGGCTACAGAGGAAATGAAGGTAC CCTAGAATCCATCAATGAGAAGACTTTTTCAGCTACTAT	378	GGGGACAAGTTTGTACAAAAA AGCAGGCTTAGAAAACCTGTA CTTCCAGTTTGAAGAGGAGAA GAAAAGAAC

					CGTCATTGAAACTGGCCCTTTAAAAGGACGCAGAGTTG AAGGAATTCAATATGAAGACATCTCTAAACTT		
KN17_G	Glycine (G)	GGC	GEEEKKRTARTDYWLQPEIIVKIITKKLGEK YHKKKAIVKEVIDKYTAVVKMIDSGDKLKL DQTHLETVIPAPGKRILVLNGGYRGNEGTL ESINEKTFSATIVIETGPLKGRRVEGIQYED ISKL	126	GGCGAAGAGGAGAAGAAAAGAACTGCCCGAACAGACT ACTGGCTACAGCCTGAAATTATTGTGAAAATTATAACCA AGAAACTGGGAGAGAAATATCATAAGAAAAAGGCTATT GTTAAGGAAGTAATTGACAAATATACAGCTGTTGTGAAG ATGATTGATTCTGGAGACAAGCTGAAACTTGACCAGAC TCATTTAGAGACAGTAATTCCAGCACCAGGAAAAAGAAT TCTAGTTTTAAATGGAGGCTACAGAGGAAATGAAGGTA CCCTAGAATCCATCAATGAGAAGACTTTTTCAGCTACTA TCGTCATTGAAACTGGCCCTTTAAAAGGACGCAGAGTT GAAGGAATTCAATATGAAGACATCTCTAAACTT	378	GGGGACAAGTTTGTACAAAAA AGCAGGCTTAGAAAACCTGTA CTTCCAGGGCGAAGAGGAGA AGAAAAGAAC
KN17_H	Histidine (H)	CAT	HEEEKKRTARTDYWLQPEIIVKIITKKLGEK YHKKKAIVKEVIDKYTAVVKMIDSGDKLKL DQTHLETVIPAPGKRILVLNGGYRGNEGTL ESINEKTFSATIVIETGPLKGRRVEGIQYED ISKL	126	CATGAAGAGGAGAAGAAAAGAACTGCCCGAACAGACTA CTGGCTACAGCCTGAAATTATTGTGAAAATTATAACCAA GAAACTGGGAGAGAAATATCATAAGAAAAAGGCTATTG TTAAGGAAGTAATTGACAAATATACAGCTGTTGTGAAGA TGATTGATTCTGGAGACAAGCTGAAACTTGACCAGACT CATTTAGAGACAGTAATTCCAGCACCAGGAAAAAGAATT CTAGTTTTAAATGGAGGCTACAGAGGAAATGAAGGTAC CCTAGAATCCATCAATGAGAAGACTTTTTCAGCTACTAT CGTCATTGAAACTGGCCCTTTAAAAGGACGCAGAGTTG AAGGAATTCAATATGAAGACATCTCTAAACTT	378	GGGGACAAGTTTGTACAAAAA AGCAGGCTTAGAAAACCTGTA CTTCCAGCATGAAGAGGAGAA GAAAAGAAC
KN17_I	Isoleucine (I)	ATT	IEEEKKRTARTDYWLQPEIIVKIITKKLGEKY HKKKAIVKEVIDKYTAVVKMIDSGDKLKLD QTHLETVIPAPGKRILVLNGGYRGNEGTLE SINEKTFSATIVIETGPLKGRRVEGIQYEDI SKL	126	ATTGAAGAGGAGAAGAAAAGAACTGCCCGAACAGACTA CTGGCTACAGCCTGAAATTATTGTGAAAATTATAACCAA GAAACTGGGAGAGAAATATCATAAGAAAAAGGCTATTG TTAAGGAAGTAATTGACAAATATACAGCTGTTGTGAAGA TGATTGATTCTGGAGACAAGCTGAAACTTGACCAGACT CATTTAGAGACAGTAATTCCAGCACCAGGAAAAAGAATT CTAGTTTTAAATGGAGGCTACAGAGGAAATGAAGGTAC CCTAGAATCCATCAATGAGAAGACTTTTTCAGCTACTAT CGTCATTGAAACTGGCCCTTTAAAAGGACGCAGAGTTG AAGGAATTCAATATGAAGACATCTCTAAACTT	378	GGGGACAAGTTTGTACAAAAA AGCAGGCTTAGAAAACCTGTA CTTCCAGATTGAAGAGGAGAA GAAAAGAAC
KN17_K	Lysine (K)	AAA	KEEEKKRTARTDYWLQPEIIVKIITKKLGEK YHKKKAIVKEVIDKYTAVVKMIDSGDKLKL DQTHLETVIPAPGKRILVLNGGYRGNEGTL ESINEKTFSATIVIETGPLKGRRVEGIQYED ISKL	126	AAAGAAGAGGAGAAGAAAAGAACTGCCCGAACAGACTA CTGGCTACAGCCTGAAATTATTGTGAAAATTATAACCAA GAAACTGGGAGAGAAATATCATAAGAAAAAGGCTATTG TTAAGGAAGTAATTGACAAATATACAGCTGTTGTGAAGA TGATTGATTCTGGAGACAAGCTGAAACTTGACCAGACT CATTTAGAGACAGTAATTCCAGCACCAGGAAAAAGAATT CTAGTTTTAAATGGAGGCTACAGAGGAAATGAAGGTAC CCTAGAATCCATCAATGAGAAGACTTTTTCAGCTACTAT CGTCATTGAAACTGGCCCTTTAAAAGGACGCAGAGTTG AAGGAATTCAATATGAAGACATCTCTAAACTT	378	GGGGACAAGTTTGTACAAAAA AGCAGGCTTAGAAAACCTGTA CTTCCAGAAAGAAGAGGAGAA GAAAAGAAC
KN17_L	Leucine (L)	CTG	LEEEKKRTARTDYWLQPEIIVKIITKKLGEK YHKKKAIVKEVIDKYTAVVKMIDSGDKLKL DQTHLETVIPAPGKRILVLNGGYRGNEGTL ESINEKTFSATIVIETGPLKGRRVEGIQYED ISKL	126	CTGGAAGAGGAGAAGAAAAGAACTGCCCGAACAGACTA CTGGCTACAGCCTGAAATTATTGTGAAAATTATAACCAA GAAACTGGGAGAGAAATATCATAAGAAAAAGGCTATTG TTAAGGAAGTAATTGACAAATATACAGCTGTTGTGAAGA TGATTGATTCTGGAGACAAGCTGAAACTTGACCAGACT CATTTAGAGACAGTAATTCCAGCACCAGGAAAAAGAATT CTAGTTTTAAATGGAGGCTACAGAGGAAATGAAGGTAC CCTAGAATCCATCAATGAGAAGACTTTTTCAGCTACTAT CGTCATTGAAACTGGCCCTTTAAAAGGACGCAGAGTTG AAGGAATTCAATATGAAGACATCTCTAAACTT	378	GGGGACAAGTTTGTACAAAAA AGCAGGCTTAGAAAACCTGTA CTTCCAGCTGGAAGAGGAGA AGAAAAGAAC
KN17_M	Metheonine (M)	ATG	MEEEKKRTARTDYWLQPEIIVKIITKKLGEK YHKKKAIVKEVIDKYTAVVKMIDSGDKLKL DQTHLETVIPAPGKRILVLNGGYRGNEGTL	126	ATGGAAGAGGAGAAGAAAAGAACTGCCCGAACAGACTA CTGGCTACAGCCTGAAATTATTGTGAAAATTATAACCAA GAAACTGGGAGAGAAATATCATAAGAAAAAGGCTATTG	378	GGGGACAAGTTTGTACAAAAA AGCAGGCTTAGAAAACCTGTA

			ESINEKTFSATIVIETGPLKGRRVEGIQYED ISKL		TTAAGGAAGTAATTGACAAATATACAGCTGTTGTGAAGA TGATTGATTCTGGAGACAAGCTGAAACTTGACCAGACT CATTTAGAGACAGTAATTCCAGCACCAGGAAAAAGAATT CTAGTTTTAAATGGAGGCTACAGAGGAAATGAAGGTAC CCTAGAATCCATCAATGAGAAGACTTTTTCAGCTACTAT CGTCATTGAAACTGGCCCTTTAAAAGGACGCAGAGTTG AAGGAATTCAATATGAAGACATCTCTAAACTT		CTTCCAGATGGAAGAGGAGAA GAAAAGAAC
KN17_N	Asparagine (N)	AAC	NEEEKKRTARTDYWLQPEIIVKIITKKLGEK YHKKKAIVKEVIDKYTAVVKMIDSGDKLKL DQTHLETVIPAPGKRILVLNGGYRGNEGTL ESINEKTFSATIVIETGPLKGRRVEGIQYED ISKL	126	AACGAAGAGGAGAAGAAAAGAACTGCCCGAACAGACTA CTGGCTACAGCCTGAAATTATTGTGAAAATTATAACCAA GAAACTGGGAGAGAAATATCATAAGAAAAAGGCTATTG TTAAGGAAGTAATTGACAAATATACAGCTGTTGTGAAGA TGATTGATTCTGGAGACAAGCTGAAACTTGACCAGACT CATTTAGAGACAGTAATTCCAGCACCAGGAAAAAGAATT CTAGTTTTAAATGGAGGCTACAGAGGAAATGAAGGTAC CCTAGAATCCATCAATGAGAAGACTTTTTCAGCTACTAT CGTCATTGAAACTGGCCCTTTAAAAGGACGCAGAGTTG AAGGAATTCAATATGAAGACATCTCTAAACTT	378	GGGGACAAGTTTGTACAAAAA AGCAGGCTTAGAAAACCTGTA CTTCCAGAACGAAGAGGAGAA GAAAAGAAC
KN17_P	Proline (P)	CCG	PEEEKKRTARTDYWLQPEIIVKIITKKLGEK YHKKKAIVKEVIDKYTAVVKMIDSGDKLKL DQTHLETVIPAPGKRILVLNGGYRGNEGTL ESINEKTFSATIVIETGPLKGRRVEGIQYED ISKL	126	CCGGAAGAGGAGAAGAAAAGAACTGCCCGAACAGACT ACTGGCTACAGCCTGAAATTATTGTGAAAATTATAACCA AGAAACTGGGAGAGAAATATCATAAGAAAAAGGCTATT GTTAAGGAAGTAATTGACAAATATACAGCTGTTGTGAAG ATGATTGATTCTGGAGACAAGCTGAAACTTGACCAGAC TCATTTAGAGACAGTAATTCCAGCACCAGGAAAAAGAAT TCTAGTTTTAAATGGAGGCTACAGAGGAAATGAAGGTA CCCTAGAATCCATCAATGAGAAGACTTTTTCAGCTACTA TCGTCATTGAAACTGGCCCTTTAAAAGGACGCAGAGTT GAAGGAATTCAATATGAAGACATCTCTAAACTT	378	GGGGACAAGTTTGTACAAAAA AGCAGGCTTAGAAAACCTGTA CTTCCAGCCGGAAGAGGAGA AGAAAAGAAC
KN17_Q	Glutamine (Q)	CAG	QEEEKKRTARTDYWLQPEIIVKIITKKLGEK YHKKKAIVKEVIDKYTAVVKMIDSGDKLKL DQTHLETVIPAPGKRILVLNGGYRGNEGTL ESINEKTFSATIVIETGPLKGRRVEGIQYED ISKL	126	CAGGAAGAGGAGAAGAAAAGAACTGCCCGAACAGACT ACTGGCTACAGCCTGAAATTATTGTGAAAATTATAACCA AGAAACTGGGAGAGAAATATCATAAGAAAAAGGCTATT GTTAAGGAAGTAATTGACAAATATACAGCTGTTGTGAAG ATGATTGATTCTGGAGACAAGCTGAAACTTGACCAGAC TCATTTAGAGACAGTAATTCCAGCACCAGGAAAAAGAAT TCTAGTTTTAAATGGAGGCTACAGAGGAAATGAAGGTA CCCTAGAATCCATCAATGAGAAGACTTTTTCAGCTACTA TCGTCATTGAAACTGGCCCTTTAAAAGGACGCAGAGTT GAAGGAATTCAATATGAAGACATCTCTAAACTT	378	GGGGACAAGTTTGTACAAAAA AGCAGGCTTAGAAAACCTGTA CTTCCAGCAGGAAGAGGAGA AGAAAAGAAC
KN17_R	Arginine (R)	CGT	REEEKKRTARTDYWLQPEIIVKIITKKLGEK YHKKKAIVKEVIDKYTAVVKMIDSGDKLKL DQTHLETVIPAPGKRILVLNGGYRGNEGTL ESINEKTFSATIVIETGPLKGRRVEGIQYED ISKL	126	CGTGAAGAGGAGAAGAAAAGAACTGCCCGAACAGACTA CTGGCTACAGCCTGAAATTATTGTGAAAATTATAACCAA GAAACTGGGAGAGAAATATCATAAGAAAAAGGCTATTG TTAAGGAAGTAATTGACAAATATACAGCTGTTGTGAAGA TGATTGATTCTGGAGACAAGCTGAAACTTGACCAGACT CATTTAGAGACAGTAATTCCAGCACCAGGAAAAAGAATT CTAGTTTTAAATGGAGGCTACAGAGGAAATGAAGGTAC CCTAGAATCCATCAATGAGAAGACTTTTTCAGCTACTAT CGTCATTGAAACTGGCCCTTTAAAAGGACGCAGAGTTG AAGGAATTCAATATGAAGACATCTCTAAACTT	378	GGGGACAAGTTTGTACAAAAA AGCAGGCTTAGAAAACCTGTA CTTCCAGCGTGAAGAGGAGA AGAAAAGAAC
KN17_S	Serine (S)	AGC	SEEEKKRTARTDYWLQPEIIVKIITKKLGEK YHKKKAIVKEVIDKYTAVVKMIDSGDKLKL DQTHLETVIPAPGKRILVLNGGYRGNEGTL ESINEKTFSATIVIETGPLKGRRVEGIQYED ISKL	126	AGCGAAGAGGAGAAGAAAAGAACTGCCCGAACAGACT ACTGGCTACAGCCTGAAATTATTGTGAAAATTATAACCA AGAAACTGGGAGAGAAATATCATAAGAAAAAGGCTATT GTTAAGGAAGTAATTGACAAATATACAGCTGTTGTGAAG ATGATTGATTCTGGAGACAAGCTGAAACTTGACCAGAC TCATTTAGAGACAGTAATTCCAGCACCAGGAAAAAGAAT TCTAGTTTTAAATGGAGGCTACAGAGGAAATGAAGGTA CCCTAGAATCCATCAATGAGAAGACTTTTTCAGCTACTA	378	GGGGACAAGTTTGTACAAAAA AGCAGGCTTAGAAAACCTGTA CTTCCAGAGCGAAGAGGAGA AGAAAAGAAC

					TCGTCATTGAAACTGGCCCTTTAAAAGGACGCAGAGTT GAAGGAATTCAATATGAAGACATCTCTAAACTT		
KN17_T	Threonine (T)	ACC	TEEEKKRTARTDYWLQPEIIVKIITKKLGEK YHKKKAIVKEVIDKYTAVVKMIDSGDKLKL DQTHLETVIPAPGKRILVLNGGYRGNEGTL ESINEKTFSATIVIETGPLKGRRVEGIQYED ISKL	126	ACCGAAGAGGAGAAGAAAAGAACTGCCCGAACAGACTA CTGGCTACAGCCTGAAATTATTGTGAAAATTATAACCAA GAAACTGGGAGAGAAATATCATAAGAAAAAGGCTATTG TTAAGGAAGTAATTGACAAATATACAGCTGTTGTGAAGA TGATTGATTCTGGAGACAAGCTGAAACTTGACCAGACT CATTTAGAGACAGTAATTCCAGCACCAGGAAAAAGAATT CTAGTTTTAAATGGAGGCTACAGAGGAAATGAAGGTAC CCTAGAATCCATCAATGAGAAGACTTTTTCAGCTACTAT CGTCATTGAAACTGGCCCTTTAAAAGGACGCAGAGTTG AAGGAATTCAATATGAAGACATCTCTAAACTT	378	GGGGACAAGTTTGTACAAAAA AGCAGGCTTAGAAAACCTGTA CTTCCAGACCGAAGAGGAGA AGAAAAGAAC
KN17_V	Valine (V)	GTT	VEEEKKRTARTDYWLQPEIIVKIITKKLGEK YHKKKAIVKEVIDKYTAVVKMIDSGDKLKL DQTHLETVIPAPGKRILVLNGGYRGNEGTL ESINEKTFSATIVIETGPLKGRRVEGIQYED ISKL	126	GTTGAAGAGGAGAAGAAAAGAACTGCCCGAACAGACTA CTGGCTACAGCCTGAAATTATTGTGAAAATTATAACCAA GAAACTGGGAGAGAAATATCATAAGAAAAAGGCTATTG TTAAGGAAGTAATTGACAAATATACAGCTGTTGTGAAGA TGATTGATTCTGGAGACAAGCTGAAACTTGACCAGACT CATTTAGAGACAGTAATTCCAGCACCAGGAAAAAGAATT CTAGTTTTAAATGGAGGCTACAGAGGAAATGAAGGTAC CCTAGAATCCATCAATGAGAAGACTTTTTCAGCTACTAT CGTCATTGAAACTGGCCCTTTAAAAGGACGCAGAGTTG AAGGAATTCAATATGAAGACATCTCTAAACTT	378	GGGGACAAGTTTGTACAAAAA AGCAGGCTTAGAAAACCTGTA CTTCCAGGTTGAAGAGGAGAA GAAAAGAAC
KN17_W	Tryptophane (W)	TGG	WEEEKKRTARTDYWLQPEIIVKIITKKLGEK YHKKKAIVKEVIDKYTAVVKMIDSGDKLKL DQTHLETVIPAPGKRILVLNGGYRGNEGTL ESINEKTFSATIVIETGPLKGRRVEGIQYED ISKL	126	TGGGAAGAGGAGAAGAAAAGAACTGCCCGAACAGACT ACTGGCTACAGCCTGAAATTATTGTGAAAATTATAACCA AGAAACTGGGAGAGAAATATCATAAGAAAAAGGCTATT GTTAAGGAAGTAATTGACAAATATACAGCTGTTGTGAAG ATGATTGATTCTGGAGACAAGCTGAAACTTGACCAGAC TCATTTAGAGACAGTAATTCCAGCACCAGGAAAAAGAAT TCTAGTTTTAAATGGAGGCTACAGAGGAAATGAAGGTA CCCTAGAATCCATCAATGAGAAGACTTTTTCAGCTACTA TCGTCATTGAAACTGGCCCTTTAAAAGGACGCAGAGTT GAAGGAATTCAATATGAAGACATCTCTAAACTT	378	GGGGACAAGTTTGTACAAAAA AGCAGGCTTAGAAAACCTGTA CTTCCAGTGGGAAGAGGAGA AGAAAAGAAC
KN17_Y	Tyrosine (Y)	TAT	YEEEKKRTARTDYWLQPEIIVKIITKKLGEK YHKKKAIVKEVIDKYTAVVKMIDSGDKLKL DQTHLETVIPAPGKRILVLNGGYRGNEGTL ESINEKTFSATIVIETGPLKGRRVEGIQYED ISKL	126	TATGAAGAGGAGAAGAAAAGAACTGCCCGAACAGACTA CTGGCTACAGCCTGAAATTATTGTGAAAATTATAACCAA GAAACTGGGAGAGAAATATCATAAGAAAAAGGCTATTG TTAAGGAAGTAATTGACAAATATACAGCTGTTGTGAAGA TGATTGATTCTGGAGACAAGCTGAAACTTGACCAGACT CATTTAGAGACAGTAATTCCAGCACCAGGAAAAAGAATT CTAGTTTTAAATGGAGGCTACAGAGGAAATGAAGGTAC CCTAGAATCCATCAATGAGAAGACTTTTTCAGCTACTAT CGTCATTGAAACTGGCCCTTTAAAAGGACGCAGAGTTG AAGGAATTCAATATGAAGACATCTCTAAACTT	378	GGGGACAAGTTTGTACAAAAA AGCAGGCTTAGAAAACCTGTA CTTCCAGTATGAAGAGGAGAA GAAAAGAAC

Table S5. 5| Codon frequency used to design the variants A, B and C of 24 optimized genes encoding venom peptides. This table was constructed using the average of codon frequencies used for variant A, B and C.

Amino Acid	Codon	Number of codons	Frequency per Amino Acid
Alanine (Ala)	GCA	41	0.24
	GCC	42	0.25
	GCG	77	0.46
	GCT	8	0.05
Arginine (Arg)	AGA	0	0.00
	AGG	0	0.00
	CGA	0	0.00
	CGC	77	0.33
	CGG	0	0.00
	CGT	157	0.67
Asparagine (Asn)	AAC	132	0.57
	AAT	99	0.43
Aspartate (Asp)	GAC	81	0.42
	GAT	111	0.58
Cysteine (Cys)	TGC	331	0.54
	TGT	281	0.46
Glutamine (Gln)	CAA	61	0.44
	CAG	77	0.56
Glutamate (Glu)	GAA	146	0.72
	GAG	58	0.28
Glycine (Gly)	GGA	0	0.00
	GGC	144	0.47
	GGG	0	0.00
	GGT	162	0.53
Histidine (His)	CAC	45	0.45
	CAT	54	0.55
Isoleucine (Ile)	ATA	0	0.00
	ATC	83	0.42
	ATT	115	0.58
Leucine (Leu)	CTA	0	0.00
	CTC	1	0.01
	CTG	122	0.78
	CTT	12	0.08
	TTA	8	0.05
	TTG	13	0.08
Lysine (Lys)	AAA	184	0.64
	AAG	104	0.36
Methionine (Met)	ATG	45	1.00
Phenylalanine (Phe)	TTC	41	0.43
	TTT	55	0.57
Proline (Pro)	CCA	46	0.18
	CCC	0	0.00
	CCG	208	0.82

	CCT	1	0.00
Serine (Ser)	AGC	133	0.47
	AGT	35	0.12
	TCA	9	0.03
	TCC	44	0.16
	TCG	37	0.13
	TCT	24	0.09
Threonine (Thr)	ACA	0	0.00
	ACC	166	0.54
	ACG	100	0.32
	ACT	43	0.14
Tryptophan (Trp)	TGG	51	1.00
Tyrosine (Tyr)	TAC	67	0.51
	TAT	65	0.49
Valine (Val)	GTA	0	0.00
	GTC	28	0.20
	GTG	57	0.41
	GTT	53	0.38

Table S5. 6| Codon usage of genes encoding high and low expresser variants (HE and LE, respectively) encoding either venom peptides or the respective fusion protein; codon frequency is represented by Fc.

Amino Acid	Codon	*Fc*, HE, peptides	*Fc*, LE, peptides	*Fc*, HE, fusion	*Fc*, LE, fusion	*Fc, E. coli*
Alanine (Ala)	GCA	0.26	0.18	0.28	0.27	0.23
	GCC	0.21	0.35	0.13	0.14	0.26
	GCG	0.41	0.44	0.36	0.37	0.33
	GCT	0.12	0.03	0.23	0.22	0.18
Arginine (Arg)	AGA	0.00	0.00	0.00	0.00	0.07
	AGG	0.00	0.00	0.17	0.17	0.04
	CGA	0.00	0.00	0.00	0.00	0.07
	CGC	0.38	0.44	0.35	0.39	0.36
	CGG	0.00	0.00	0.00	0.00	0.11
	CGT	0.63	0.56	0.48	0.45	0.36
Asparagine (Asn)	AAC	0.55	0.64	0.47	0.49	0.51
	AAT	0.45	0.36	0.53	0.51	0.49
Aspartate (Asp)	GAC	0.39	0.36	0.41	0.41	0.37
	GAT	0.61	0.64	0.59	0.59	0.63
Cysteine (Cys)	TGC	0.49	0.58	0.49	0.56	0.54
	TGT	0.51	0.42	0.51	0.44	0.46
Glutamine (Gln)	CAA	0.48	0.39	0.37	0.36	0.34
	CAG	0.52	0.61	0.63	0.64	0.66
Glutamate (Glu)	GAA	0.84	0.76	0.62	0.60	0.68
	GAG	0.16	0.24	0.38	0.40	0.32
Glycine (Gly)	GGA	0.00	0.00	0.07	0.07	0.13
	GGC	0.48	0.47	0.44	0.44	0.37
	GGG	0.00	0.00	0.15	0.15	0.15
	GGT	0.52	0.53	0.33	0.34	0.35
Histidine (His)	CAC	0.26	0.48	0.33	0.35	0.43
	CAT	0.74	0.52	0.67	0.65	0.57
Isoleucine (Ile)	ATA	0.00	0.00	0.00	0.00	0.11
	ATC	0.49	0.32	0.57	0.54	0.39
	ATT	0.51	0.68	0.43	0.46	0.49
Leucine (Leu)	CTA	0.00	0.00	0.00	0.00	0.04
	CTC	0.00	0.03	0.05	0.05	0.10
	CTG	0.88	0.75	0.4	0.39	0.47
	CTT	0.06	0.09	0.23	0.24	0.12
	TTA	0.03	0.09	0.14	0.14	0.14
	TTG	0.03	0.03	0.18	0.18	0.13
Lysine (Lys)	AAA	0.65	0.65	0.80	0.80	0.74
	AAG	0.35	0.35	0.20	0.20	0.26
Methionine (Met)	ATG	1.00	1.00	1.00	1.00	1.00
Phenylalanine (Phe)	TTC	0.28	0.44	0.14	0.16	0.42
	TTT	0.72	0.56	0.86	0.84	0.58
Proline (Pro)	CCA	0.17	0.17	0.17	0.17	0.20
	CCC	0.00	0.00	0.06	0.06	0.13
	CCG	0.82	0.83	0.64	0.64	0.49

	CCT	0.02	0.00	0.13	0.13	0.18
Serine (Ser)	AGC	0.51	0.40	0.67	0.65	0.25
	AGT	0.09	0.17	0.11	0.13	0.16
	TCA	0.05	0.05	0.11	0.11	0.14
	TCC	0.11	0.16	0.07	0.08	0.17
	TCG	0.13	0.15	0.02	0.02	0.14
	TCT	0.11	0.07	0.02	0.01	0.15
Threonine (Thr)	ACA	0.00	0.00	0.11	0.11	0.17
	ACC	0.55	0.55	0.45	0.45	0.4
	ACG	0.29	0.26	0.17	0.17	0.25
	ACT	0.16	0.19	0.26	0.27	0.19
Tryptophan (Trp)	TGG	1.00	1.00	1.00	1.00	1.00
Tyrosine (Tyr)	TAC	0.45	0.4	0.6	0.59	0.41
	TAT	0.55	0.6	0.4	0.41	0.59
Valine (Val)	GTA	0.00	0.00	0.05	0.05	0.17
	GTC	0.23	0.23	0.27	0.27	0.20
	GTG	0.36	0.36	0.34	0.34	0.35
	GTT	0.41	0.41	0.34	0.34	0.28

Supplemental information – Chapter 6

Table S6. 1| Animals groups and number of species within each animal group used for the selection of the 4992 peptides produced recombinantly within this study.

Venomous animal	Animal family	Number of species
Snakes	*Elapidae*, *Viperidae*, *Atractaspididae*	38
Scorpions	*Buthidae*, *Euscorpiidae*, *Hemiscorpiidae*, *Scorpionidae*, *Scorpionidae*, *Vaejovidae*	41
Cone snails	*Conidae*	40
Spiders	*Agelenidae, Araneidae,Ctenidae , Dipluridae, Lycosidae, Nephilidae,Sparassidae, Theraphosidae, Theridiidae*	46
Fishes	*Potamotrygonidae*, *Plotosidae*, *Siluridae*, *Synanceiidae*	9
Hymenoptera	*Apidae*, *Vespidae*	8
Scolopendra	*Scolopendridae*	5
Terebra/mitre	*Mitridae*, *Terebridae*	4
Cnidarians	*Actiniidae*, *Cassiopeidae, Stichodactylidae*	3
Octopus	*Octopodidae*	3
Ants	*Formicidae*, *Myrmicinae*	2
Rays	*Potamotrygonidae*	1
Lizards	*Helodermatidae*	1
Total		201

Table S6. 2| Codon frequency used to design 4992 optimized genes encoding venom peptides.

Amino Acid	Codon	Frequency per Amino Acid
Alanine (Ala)	GCA	0.28
	GCC	0.32
	GCG	0.40
	GCT	0.00
Arginine (Arg)	AGA	0.00
	AGG	0.00
	CGA	0.00
	CGC	0.50
	CGG	0.00
	CGT	0.50
Asparagine (Asn)	AAC	0.51
	AAT	0.49
Aspartate (Asp)	GAC	0.37
	GAT	0.63
Cysteine (Cys)	TGC	0.50
	TGT	0.50
Glutamine (Gln)	CAA	0.34
	CAG	0.66
Glutamate (Glu)	GAA	0.68
	GAG	0.32
Glycine (Gly)	GGA	0.00
	GGC	0.52
	GGG	0.00
	GGT	0.48
Histidine (His)	CAC	0.43
	CAT	0.57
Isoleucine (Ile)	ATA	0.00
	ATC	0.44
	ATT	0.56
Leucine (Leu)	CTA	0.00
	CTC	0.10
	CTG	0.50
	CTT	0.12
	TTA	0.15
	TTG	0.13
Lysine (Lys)	AAA	0.74
	AAG	0.26
Methionine (Met)	ATG	1.00

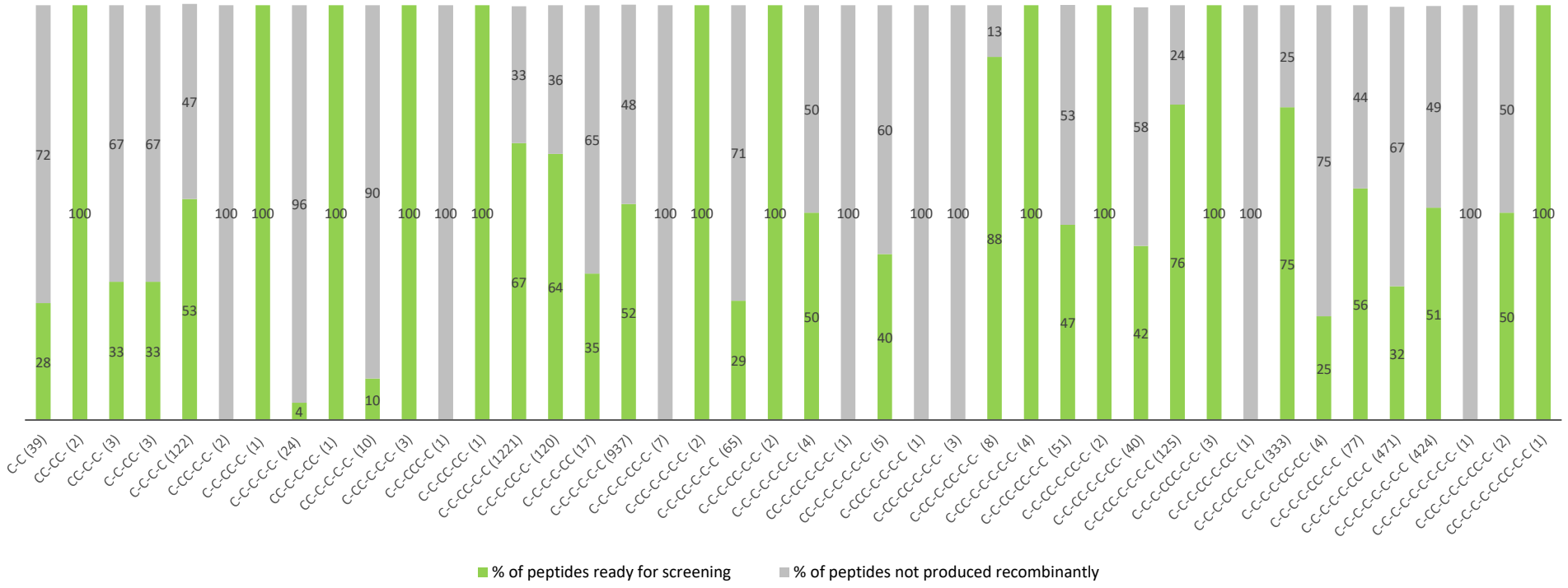
Phenylalanine (Phe)	TTC	0.42
	TTT	0.58
Proline (Pro)	CCA	0.29
	CCC	0.00
	CCG	0.71
	CCT	0.00
Serine (Ser)	AGC	0.36
	AGT	0.23
	TCA	0.00
	TCC	0.21
	TCG	0.20
	TCT	0.00
Threonine (Thr)	ACA	0.00
	ACC	0.48
	ACG	0.30
	ACT	0.22
Tryptophan (Trp)	TGG	1.00
Tyrosine (Tyr)	TAC	0.41
	TAT	0.59
Valine (Val)	GTA	0.00
	GTC	0.24
	GTG	0.42
	GTT	0.34

Table S6. 3| Principal characteristics of 4963 venom peptides and expression yields obtained in *E. coli*.

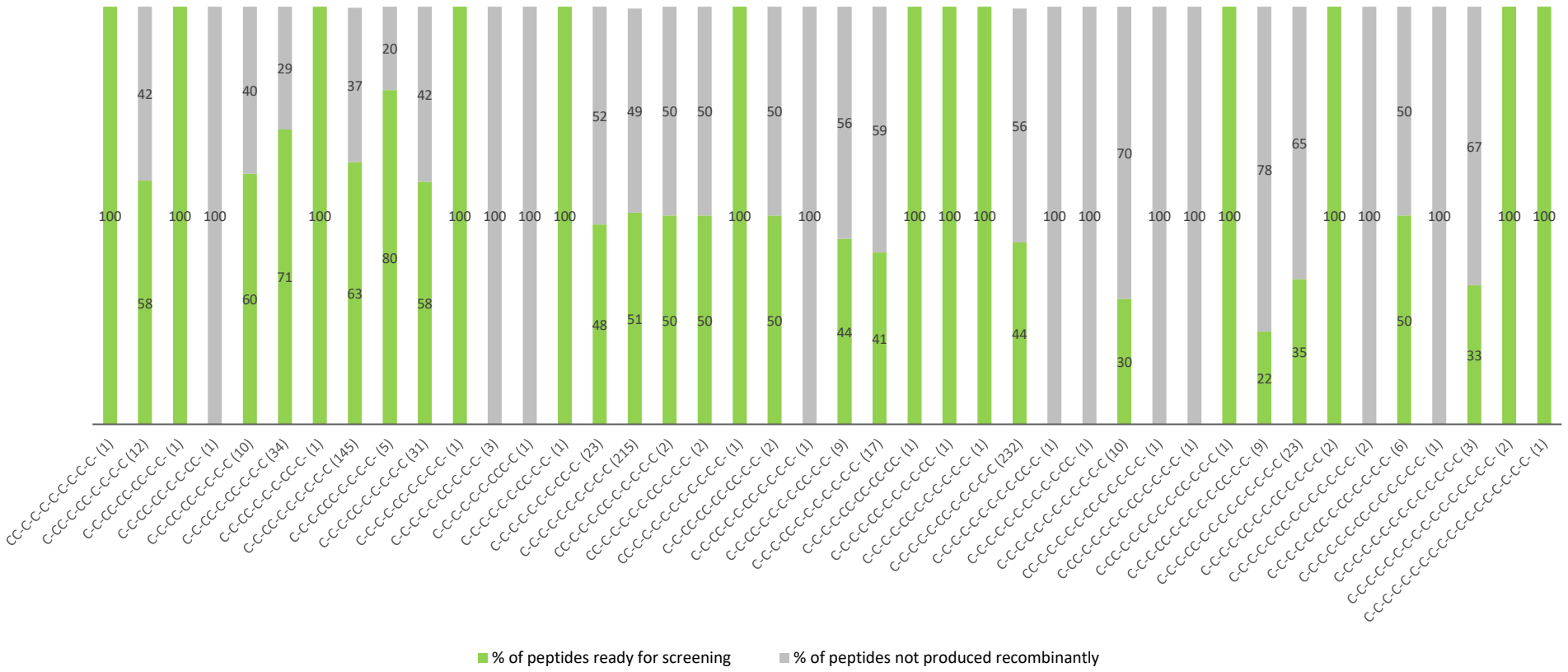
This table is available in digital format.

Figure S6. 1| Influence of cysteine pattern in the capacity of *E. coli* to produce animal venom peptides. The 4992 venom peptides produced in this project represent 84 different cysteine patterns (Panel A: first 42 patterns; panel B: second group of 42 patterns). The figure correlates the type of cysteine pattern with percentage of peptides efficiently produced in *E. coli* (green) or not efficiently produced in the bacterial host (gray).

A.



B.


% of peptides ready for screening
% of peptides not produced recombinantly
MM